Fantasy Football:

"How to get a leg up"

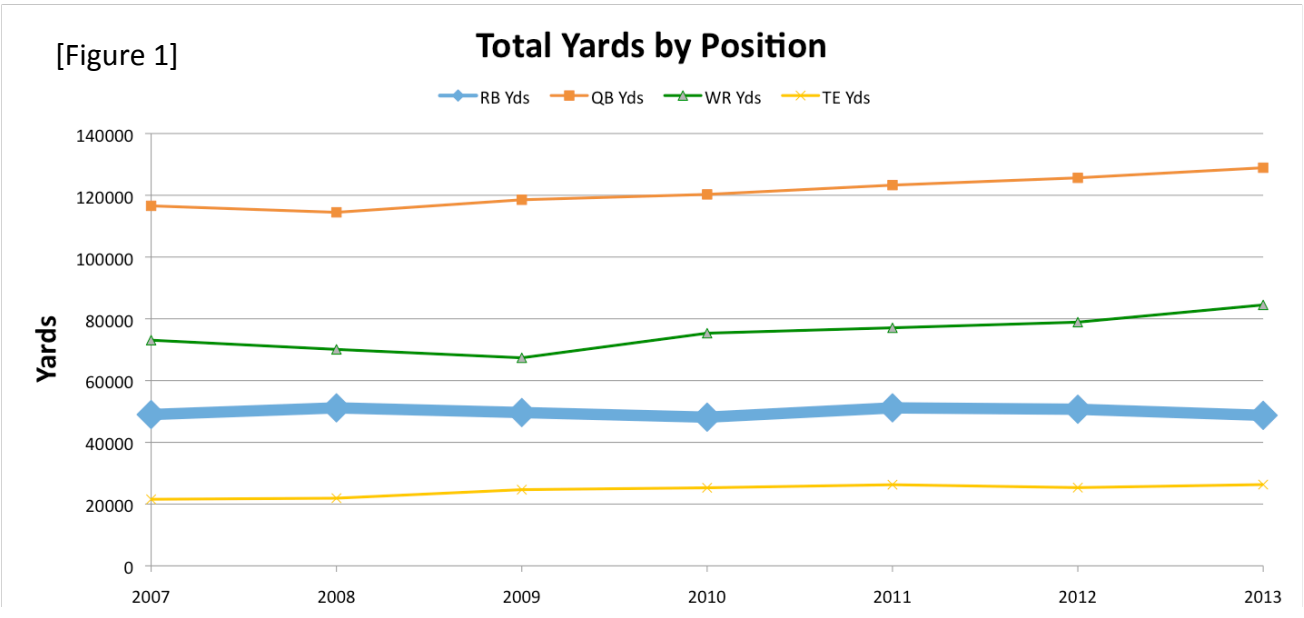
By: Chris Balthazard, Troy Holland, Ching-Hao Hu, Nik Oza, and Xavier Weisenreder

Fantasy football is an online game where each player in his or her league acquires a set of players to formulate a team, in which those players' NFL statistics correspond to fantasy points. Team owners select their roster's players prior to the season in a draft. Additionally, throughout the season, players have the ability to pick up "free agents," players not on other teams. Each week, there are matchups between fantasy teams, and the team with the highest point total from all of his or her players wins the matchup. Similar to the NFL, teams with good records make the fantasy playoffs, which crown a champion near the end of the NFL regular season.

Just as in real football, there exists a lot of variation in fantasy football. No expert or prospective player can predict with 100% certainty the outcome of any given season; however, every expert and prospective player *will* assert that it is possible to optimize one's chances for fantasy success by understanding certain NFL trends. In this series of data science analyses, we examine a few of those trends and, in the end, make an informed conclusion about how to approach fantasy football.

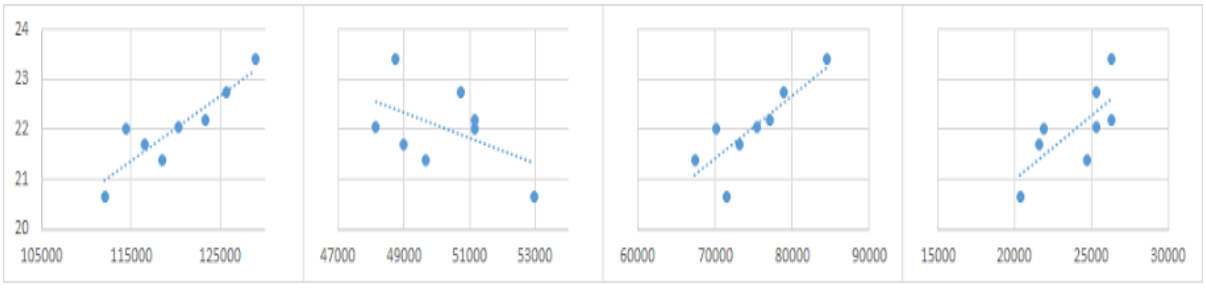
Our first analysis examines whether or not it is plausible to confidently utilize a series of yearly statistical trends in an effort to gain an edge in fantasy football. This examination begins with an analysis of yearly positional yardage totals and ends with the introduction of a fantasy football specific statistic called value-based drafting (VBD). Individual player data was collected from Pro-Football-Reference for the years 2007 to 2013 and included 3,899 unique player entries. Using data outliers as a measurement of its cleanliness, the collected data was extremely clean. There did exist a high amount of missing values in the data; however, it was quickly noticed that the site tended to use empty data attributes as a symbol for a value of zero, and we found the fix to be trivial. In general, the collected data was relatively easy to understand, manipulate, and fix when necessary, which ultimately made drawing conclusions much easier.

[Figure 1]



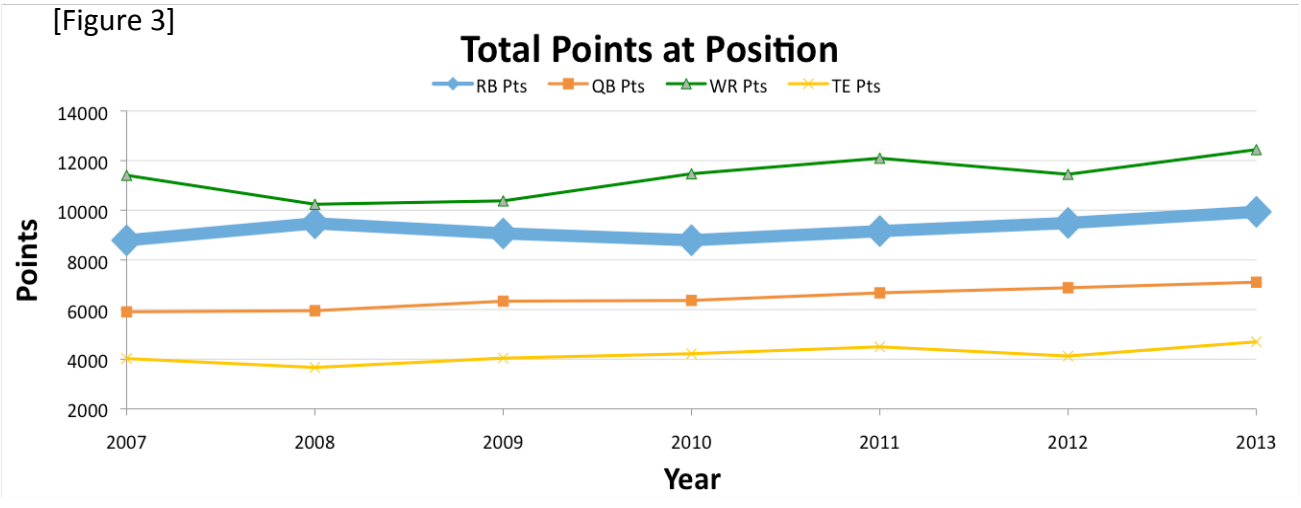
We begin with an examination of yearly total positional yardage for quarterbacks (QB), running backs (RB), wide receivers (WR), and tight ends (TE) from 2007 to 2013. From the collected data, we separately pooled all passing yards for QBs, rushing yards for RBs, and receiving yards for WRs and TEs for each year from 2007 to 2013 and plotted the results on an x-y scatter plot (Figure 1). The results were relatively expected. That is, passing-related statistics have increased, as QB passing yards, WR receiving yards, and TE receiving yards are all trending upwards. On the contrary, RB rushing yards are trending slightly downwards, and are certainly not increasing at the rate of other yards. In short, the NFL is becoming more passing-oriented as passing has undertaken a greater role in offenses and has done so at the expense of RBs.

[Figure 2]



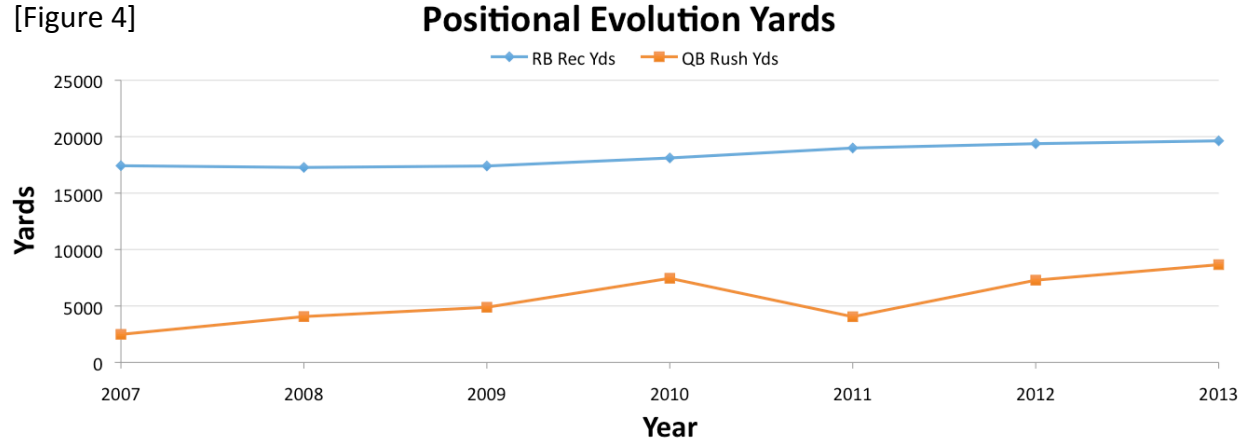
From these data points used on the x-y scatter plot, we created a scatter plot matrix (Figure 2) to show all the correlations between passing yards for all QBs, rushing yards for all RBs, and receiving yards for all WRs and TEs in the years 2007 to 2013, where each point represents the total yards in the category for the entire season. Additionally, we plotted the correlations against average score per game (from Professor Singh's given data) for the seasons 2007 to 2013, because the score data only went back to 2007. Unfortunately, the data only covers 7 (or 8 for score) seasons, which leads to few data points with which to find trends, but with those points, the results that we get are relatively clear: looking inside the boxed area, the increases in QB passing yards and receiving yards for both WRs and TEs has correlated with the increase in average game score in the past few seasons. The RB-score chart is more varied and the result slightly less clear, but the trendline displays a downward trend, which makes sense given the rest of the RB correlations. So if the league has become more passing-oriented over this time period, what do these trends mean in terms of fantasy?

[Figure 3]



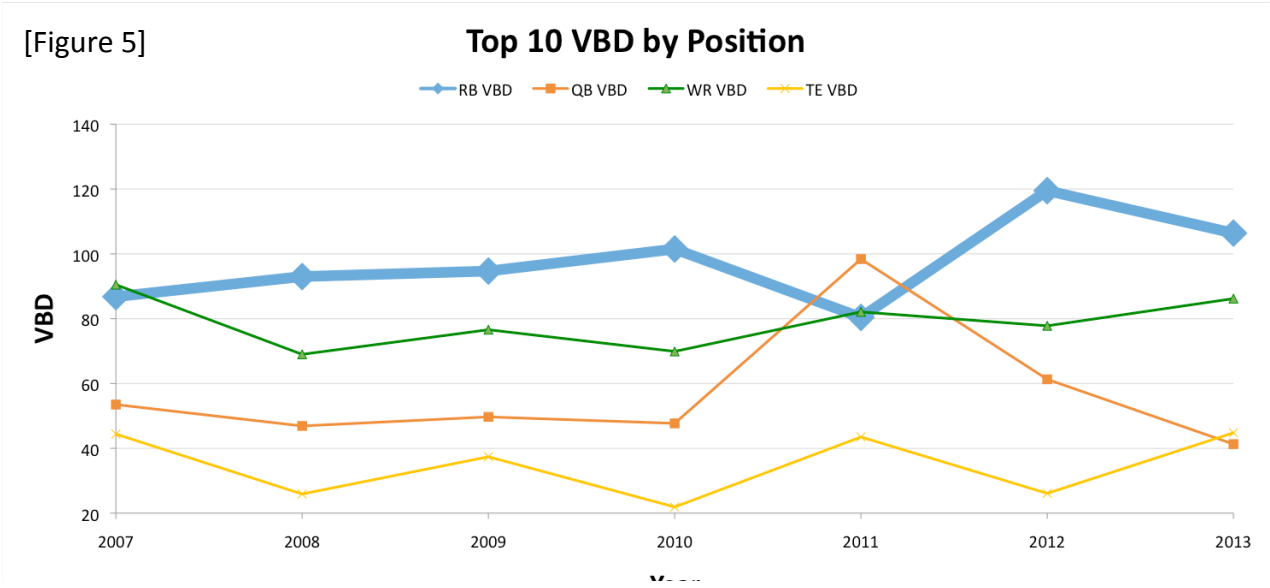
We wanted to figure out whether those yearly yardage trends for each position successfully mapped to yearly fantasy points trends as well. We pooled fantasy points and plotted results on an x-y scatter plot (Figure 3). The findings demonstrate that during this time period, passing-related positions (QB, WR, and TE) have shown an increase in total fantasy production; however, in spite of the fact that rushing yardage totals for RBs have been shown to decrease over this time period, the total fantasy production of the RB position has remained relatively constant. Although the NFL has become an increasingly more passing-oriented league, RBs have continued to be incorporated into NFL offenses; however, they have increasingly done so by means of the passing game.

[Figure 4]



In an effort to examine the evolution of this position, we pooled receiving yards for RBs and plotted them in an x-y scatter plot (Figure 4). The results highlight that although the rushing yardage totals of RBs have slightly decreased from 2007 to 2013, RBs have been progressively swapping rushing opportunities with receiving opportunities in modern NFL offenses, which further explains their stable fantasy production during this time period. Consequently, we hypothesized that NFL offenses have compensated for this lack of rushing from RBs by turning to QBs for rushing production. After plotting the rushing yards of QBs during this time period, our hypothesis was confirmed. In the same time period, total QB rushing yards have increased. So have these positional evolutions, along with the fact that the league has gradually become more passing-oriented, altered the importance of certain positions in fantasy? That is, are QBs now more valuable and RBs less valuable to fantasy football than in previous years?

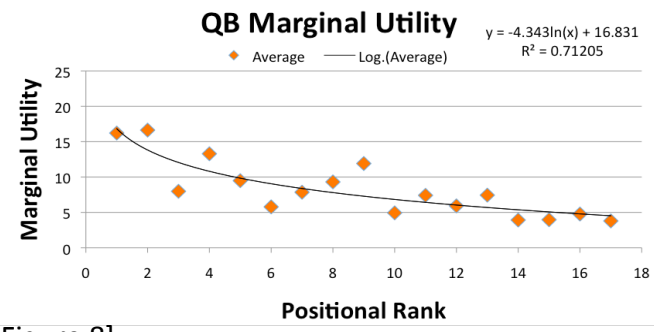
[Figure 5]



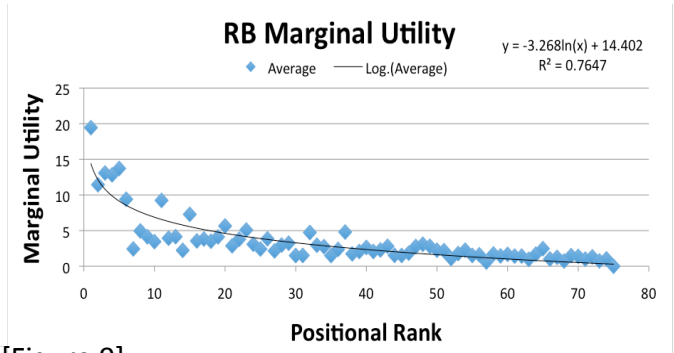
To answer that question, we examined VBD for each position from 2007 to 2013 with the same players as above. In short, VBD is a measure of how much better a player is compared to the baseline player at his position. At each position, we took the average of the top ten players' VBD for each year and graphed the results in an x-y scatter plot (Figure 5). First of all, we must note 2011, a year in which four QBs (Drew Brees, Tom Brady, Matthew Stafford, and Eli Manning) produced record breaking statistics that greatly increased the VBD of the top ten players at the QB position. Aside from that outlier in the results, we found that, on average, the top ten RBs consistently have the highest VBD. The list rounds out with WRs, QBs, and TEs. Each year, therefore, the top 10 RBs are *most important* to my fantasy football team. That is, as fantasy football players, we want a top 10 RB on our teams more than we want a top 10 player from any other position.

We also found other VBD positional trends when examining the data. From 2007 to 2013, the VBD of the top ten players at passing-related positions (QBs, WR, and TEs) has remained relatively constant; however, the VBD of the top ten RBs has increased during that same time period. Effectively, a top ten RB, nowadays, is *more important* to a fantasy football team than a top ten RB was in 2007. The same, however, cannot be said for passing-related positions. Although statistics show that from 2007 to 2013 the NFL has become a more passing-oriented league, this does not mean that we should now view the top ten players at passing-related positions as more important to a fantasy football team than we did in 2007. Rather, the results suggest the opposite. We want as many top ten RBs on our fantasy football teams as possible, and we must implement that strategy in the fantasy draft.

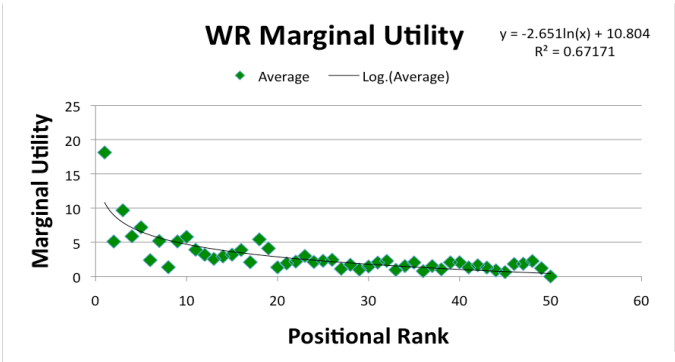
[Figure 6]



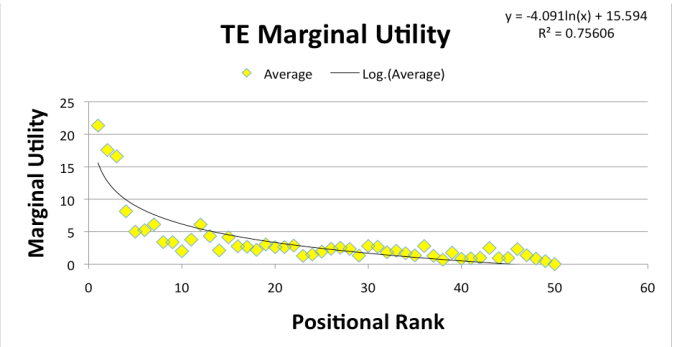
[Figure 7]



[Figure 8]

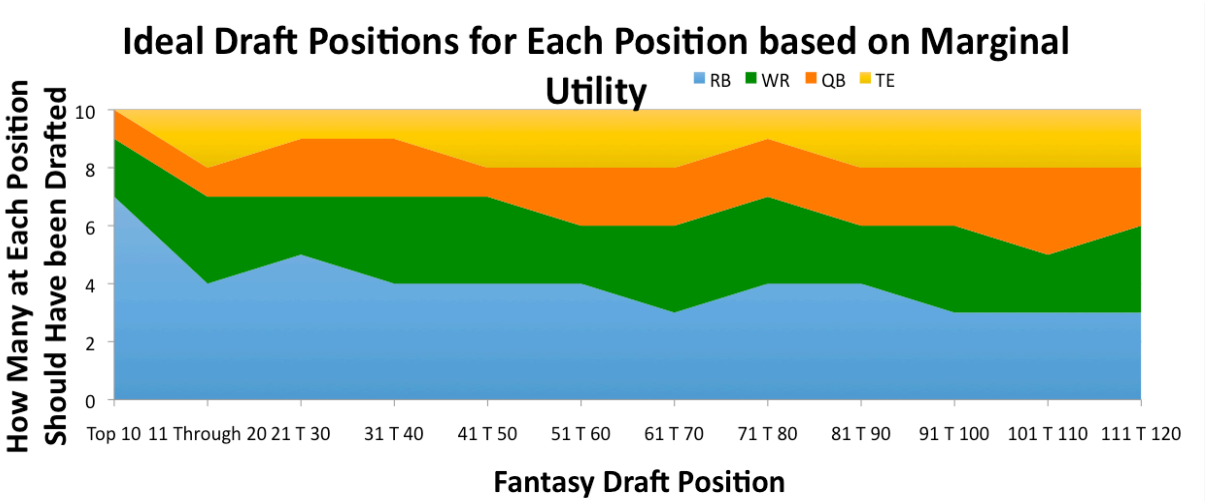


[Figure 9]



We then looked at how optimal drafting is done based on the positions that are drafted. We can look at the marginal utility for each position over the last 6 years, which is the average difference between the top ranked player at that position, and the 2nd ranked player, the difference between the 2nd and the 3rd, etc. Only the top ranks were used for each position (top 16 for QBs, 75 for RBs, 50 for WRs and 50 for TEs). As we can see for each of the plots of marginal utility vs. rank (Figure 6, 7, 8, 9), a logarithmic fit appears to fit best for each plot (and makes intuitive sense given knowledge of the sport). Having established these curves, we can then plug the ranks of the positions into the logarithmic function so that we get an estimate for the marginal utility for each rank of every position. We had to apply some adjustments to the estimated marginal utility, given the setup of fantasy football (that is, the QB value is halved because there is only one QB slot, and the TE value is multiplied by 2/3).

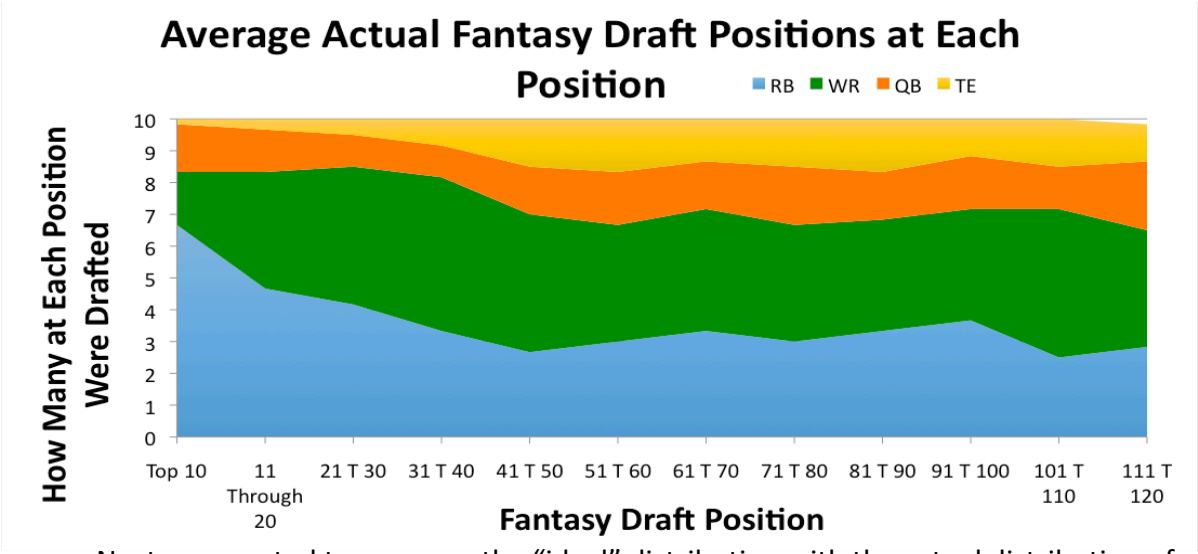
[Figure 10]



Having established these marginal utility estimates for each rank at each position, we were able to combine these ranks to produce our ranked list of marginal utility, that is the highest marginal utility at a given draft position regardless of position. From here we were able to establish a projection for where positions should ideally go in the draft in order to maximize marginal utility (Figure 10).

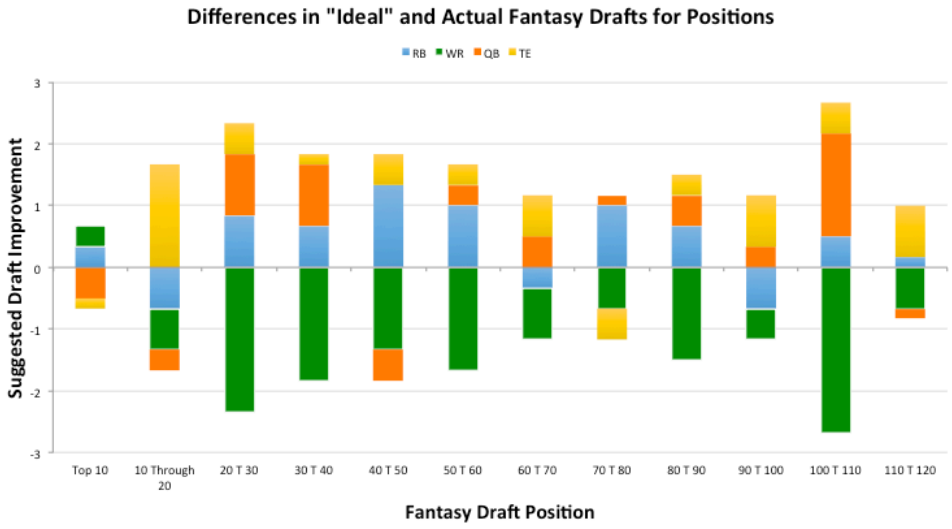
From here, we wanted to compare our ideal model with empirical fantasy draft data, where each position was actually being drafted in fantasy drafts. For this, we looked at the average draft positions over the last 10 years, for each of the top ranked players at each position, and then combined these drafted rank averages into one ranking, with the positions weaving in and out based on their overall average draft position. From here we can make the actual distribution of where players have gone on average in actual fantasy drafts (Figure 11).

[Figure 11]



Next, we wanted to compare the "ideal" distribution with the actual distribution of drafted players over the last 6 years in order to find where inefficiencies exist in drafting in terms of position. For this, we simply subtracted the actual distribution from the ideal distribution in order to get an efficiency distribution that shows which positions should have been drafted more or less by round (Figure 12). We can see some trends, including general over-drafting of WRs and slight under-drafting of RBs in later rounds. On the whole, however, the drafting appears to be relatively efficient, and there are not any huge improvements that will greatly overhaul conventional drafting strategy. Clearly, this is just a quick snapshot of drafting and other local considerations (e.g. what positions you have already drafted, the talent for that year at each position) should also be taken into account, but the results are definitely interesting. Our results illustrate that fantasy drafting has been relatively efficient, but that certain inefficiencies (i.e. over-drafting of WRs) exist and can be exploited.

[Figure 12]



One of the other local considerations mentioned above is the weather of NFL games. It seems intuitive to believe that weather plays a huge role in the fantasy performance of different positions in the NFL. When we see a quarterback struggle mightily in a snowstorm, it is easy to jump to the conclusion that weather has a significant impact on the fantasy performance of players. When deciding the players to start for a particular week, weather seems to be a reasonable variable to take into account. Another variable to take into consideration when making such decisions is the opponents the players are facing. Is weather as significant to this decision as opponent defenses? The focus here is on quarterbacks and running backs, though kickers presumably could be strongly affected by weather as teams rarely attempt field goals or even kicks for the extra point in extreme weather conditions.

Data on weather information were gathered from Pro Football Reference, which has weather information available for about 75% of the games from 2006 to 2013. This includes the temperature (°F), the relative humidity (%) and the wind speed (mph) of the game.

To get a sense of the direction and strength of the effect of temperature, relative humidity and wind speed on fantasy scores of QB and RB, the correlation between the weather information and the fantasy score of QB and RB is computed.

	QB Fantasy Points	RB Fantasy Points
Temperature	0.0133	-0.0238
Relative Humidity	-0.0406	-0.0118
Wind Speed	-0.0561	0.00308

There is very slight correlation between fantasy points and temperature, humidity or wind speed for RBs. The correlation is much stronger for QBs, especially for humidity and wind speed. However, the strength of the relationship can be described as moderate at best. Contrast this with the correlation of fantasy points and measures of opponent defenses for QBs and RBs. Opponent defenses are measured as their seasonal passing/rushing yards allowed, passing/rushing touchdowns allowed and interceptions. These are adjusted to the baseline of league average so that changes in the league over time do not affect the results.

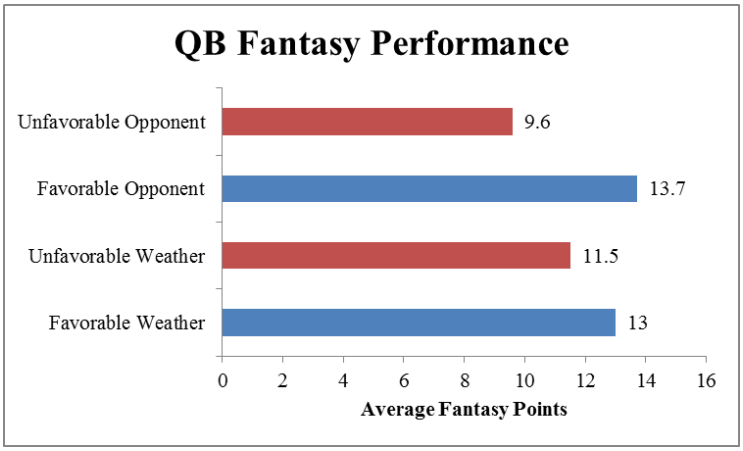
	QB Fantasy Points	RB Fantasy Points
Passing/Rushing Yards allowed adjusted	0.148	0.112
Passing/Rushing Touchdowns allowed adjusted	0.158	0.110
Interceptions adjusted	-0.0838	

The correlation of fantasy points and measures of opponent defenses is much higher, suggesting that opponent defenses play a larger role in the variance of weekly fantasy points of QBs and RBs than weather.

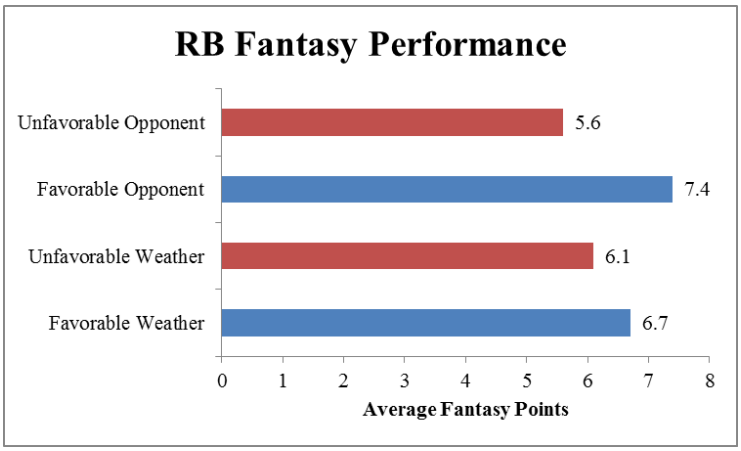
To visualize the difference between the effects of weather and opponent defenses on weekly fantasy performance, we aggregate the games that are most and least conducive for QB and RB performance in terms of weather conditions and opponent defenses. For QBs, favorable weather conditions include above-average temperature, below-average humidity and below-

average wind speed. For RBs, favorable weather conditions include below-average temperature, below-average humidity and above-average wind speed. For QBs, a favorable opponent includes a defense with above-average passing yards allowed, above-average passing touchdowns and below-average interceptions. For RBs, a favorable opponent includes a defense with above-average passing yards allowed, above-average passing touchdowns and below-average interceptions. The average player performances for QBs and RBs in these four conditions are displayed in Figures 13 and 14, respectively.

[Figure 13]



[Figure 14]



The gap between QB and RB fantasy performances in favorable and unfavorable weather conditions is much smaller than the gap between QB and RB fantasy performances against favorable and unfavorable opponents. This further supports the point shown with the correlations that weather is not as crucial a factor as an opponent's strength when it comes to impacting weekly fantasy performance of QB and RB.

To strip out the effects of random variation and other factors in play, we performed a fixed effects regression that also controls for the surface and location (home/away) as well as

weather conditions and measures of opponent defense. The fixed effects regression also controls for the identity of the players in the games, so the regression is essentially comparing the performance of the same player in different conditions.

The results of the fixed effects regression further support the previous results that weather plays a significantly smaller role than opponent defense in the variation of fantasy performance of QBs and RBs. For the regression on the QB data, only wind speed shows up to be significant in the weekly fantasy points of the three weather elements, while temperature and humidity appear highly insignificant, with p-values larger than 0.5. Meanwhile, all three aspects of the passing defense are highly significant with p-values lower than 0.001. The magnitudes of the effects of opponent defense are stronger as well. A pass defense 10% worse than average in every aspect is expected to raise the QB's fantasy points by 0.64 + 0.49 + 0.15 = 1.28 points. On the other hand, an increase of 10 mph in wind speed, a significant change, is only expected to lower the QB's fantasy points by 0.73 points. For reference, playing at home raises a QB fantasy points by 0.64 points versus playing in an away game, on average.

For the regression on the RB data, only temperature shows up to be highly significant in the weekly fantasy points of the three weather elements while humidity appears to be moderately significant with a p-value around 0.05. Temperature appears not to be significant at all with a p-value larger than 0.5. Meanwhile, both aspects of the rushing defense are highly significant with p-values lower than 0.001. The magnitudes of the effects of opponent defense are stronger as well. A rush defense 10% worse than average in every aspect is expected to raise the RB's fantasy points by 0.26 + 0.12 = 0.38 points. On the other hand, an increase of 20°F in temperature, a significant change, is only expected to lower the RB's fantasy points by 0.34 points. For reference, playing at home raises RB fantasy points by 0.39 points versus playing in an away game, on average.

As can be seen from the analysis of weather and opponent defense, opponent defenses play a significantly more crucial role than weather conditions in impacting the weekly fantasy performances of QB and RB. While we should not completely ignore the weather when we decide on our starters for the week, it should be a minor point of consideration compared to opponent defenses.

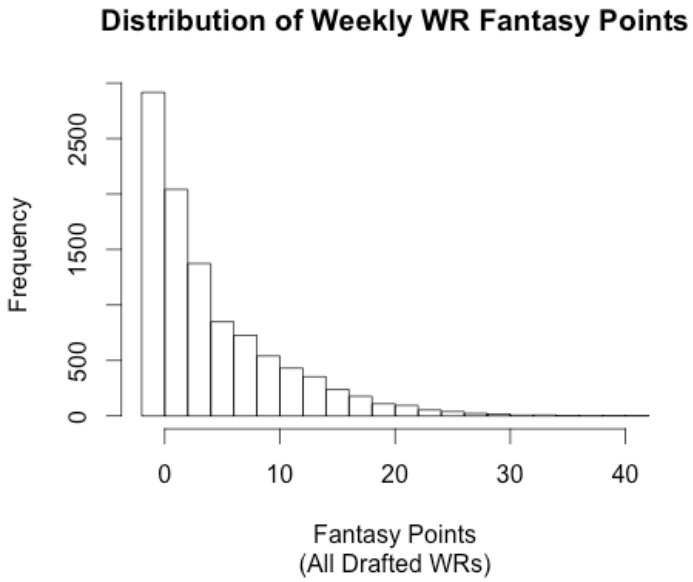
Finally, to get a better sense of which factors a fantasy football player may want to consider when drafting his or her team, we decided to look at the actual NFL draft and its effects on fantasy production. We collected draft data from Pro Football Reference from the 2004 to the 2013 drafts. Because many players played in the league during that timeframe but were drafted before 2004, they were excluded from the following analyses.

Additionally, fullbacks and those players listed as fullback-TEs were excluded from the RB subset of the dataset because these players usually do not get many rushing attempts and and therefore mostly irrelevant to fantasy football. Leaving them in the dataset might paint a misleading picture about the distribution of RB fantasy points. Fullback-TEs were also excluded from the TE subset because those players also do not get many receiving targets and are therefore irrelevant to fantasy football.

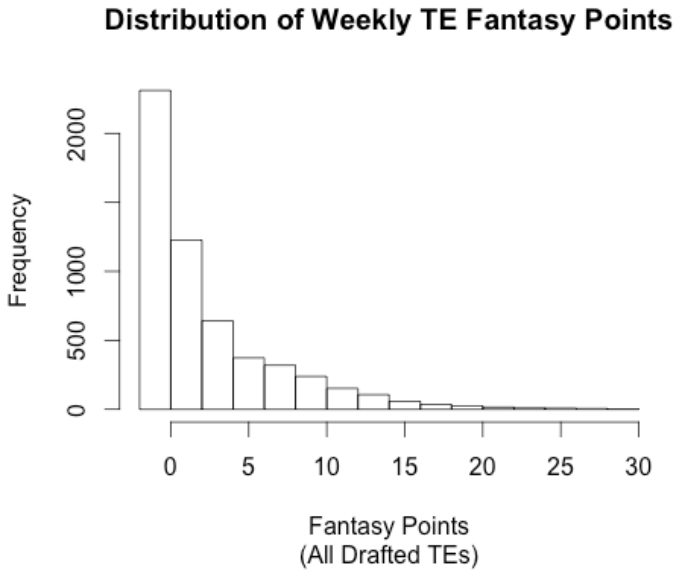
For exploratory analysis, we graphed histograms of fantasy points and split by position to get a sense of the overall distributions of fantasy points. Not surprisingly, for each position, the distribution of fantasy points is very skewed to the right (Figures 15-18). This is because fantasy points (and really all traditional football stats) are extremely playing-time dependent.

However, it would not make sense to remove extreme values because they reflect this part of fantasy football.

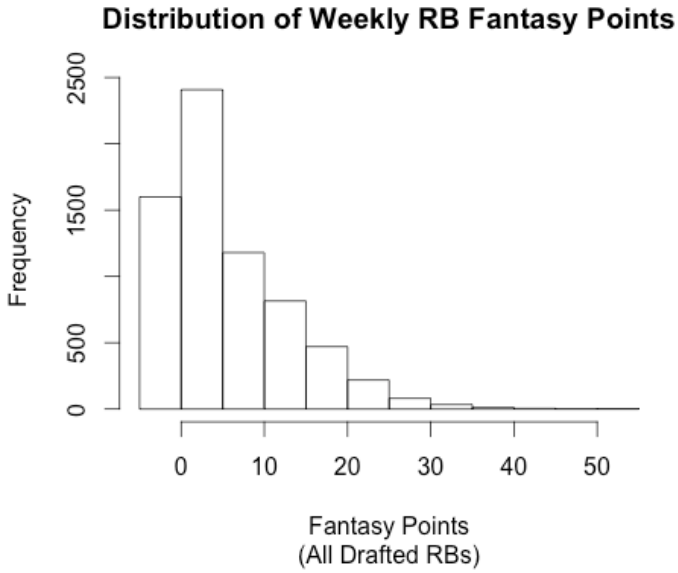
[Figure 15]



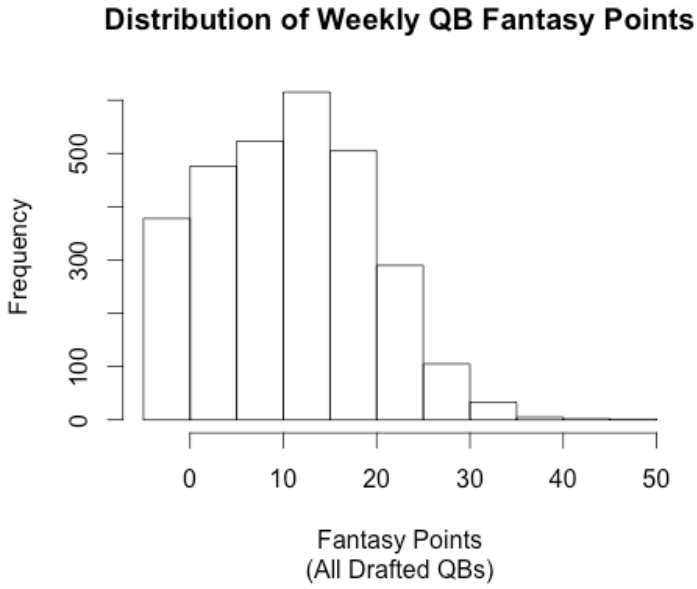
[Figure 16]



[Figure 17]

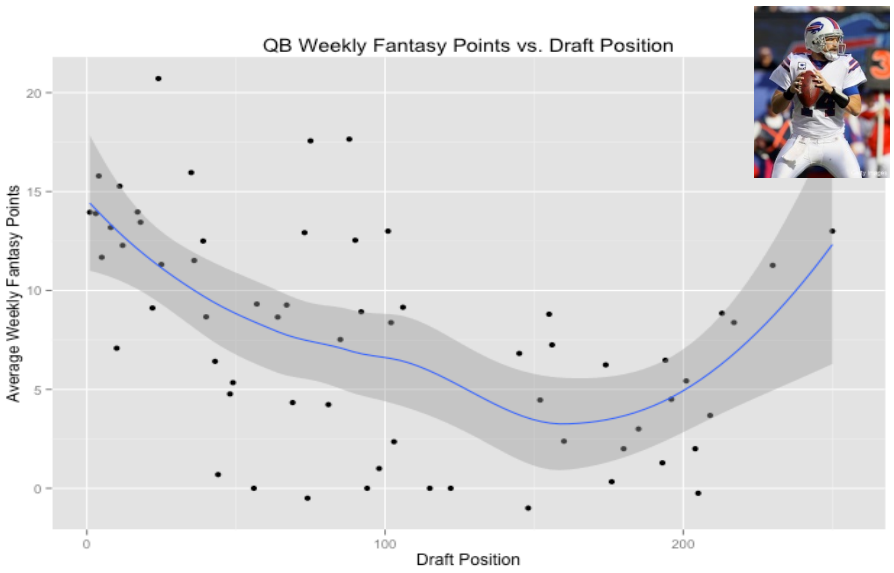


[Figure 18]

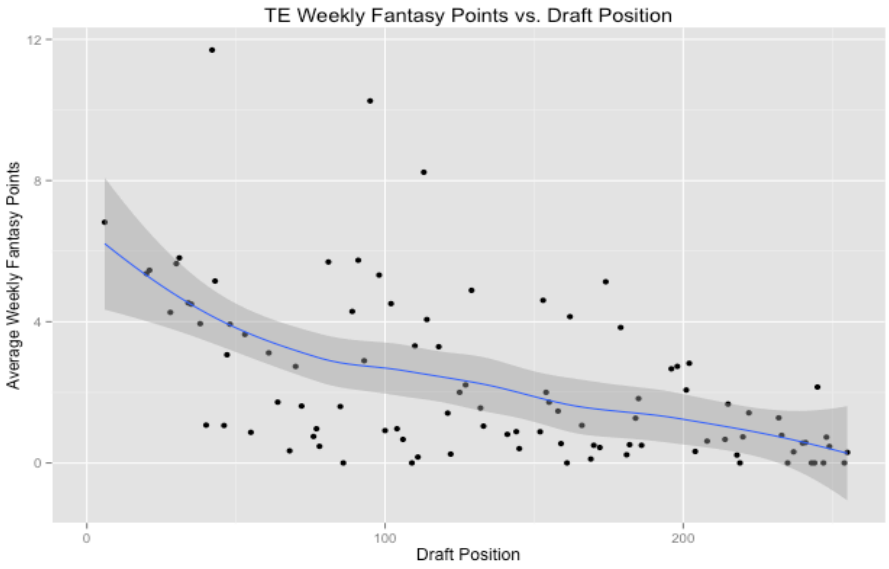


Moving on to correlations, we can see that draft position is negatively correlated with weekly fantasy points for each position. This is hardly surprising, as better players get picked higher in the draft; this trend is easily visualized by plotting average fantasy points against draft position (Figures 19-22). Despite a clear negative trend, it is important to observe the high variance for each position.

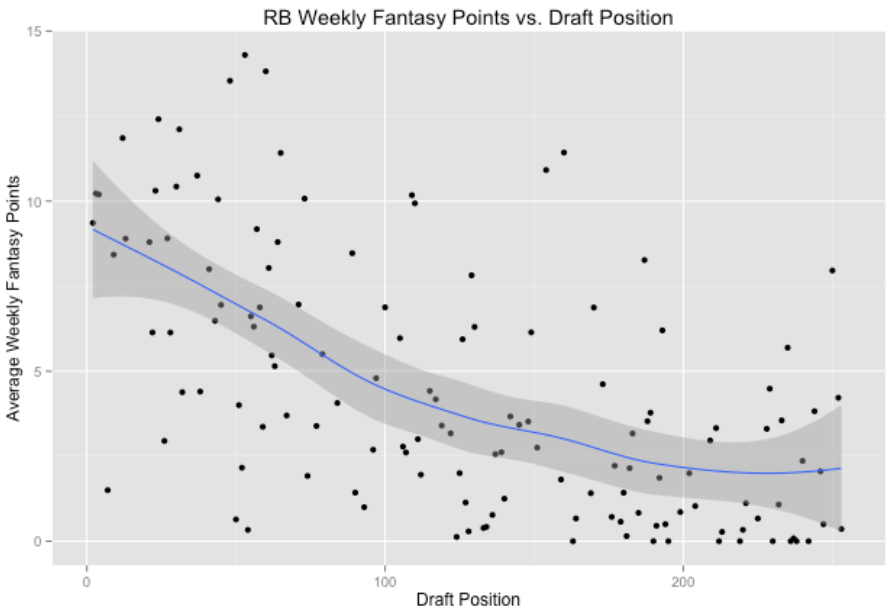
[Figure 19]



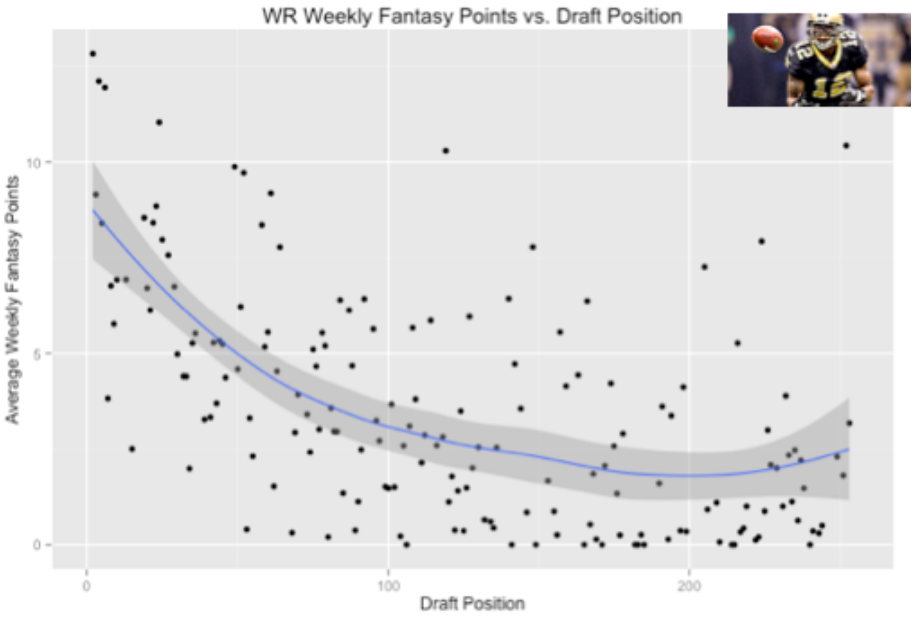
[Figure 20]



[Figure 21]



[Figure 22]



Finally, it is interesting to note two clear outliers: QB Ryan Fitzpatrick (Figure 19) and WR Marques Colston (Figure 22). Fitzpatrick has been a very solid starting QB in the NFL for the past few seasons despite being drafted in the seventh round of the NFL draft. Colston has been one of the best fantasy WRs during the past few seasons despite also being drafted in the seventh round. Both Colston and Fitzpatrick were drafted out of lesser college football programs than most other NFL players; Fitzpatrick played at Harvard University and Colston played at Hofstra University, then a Division 1-AA program (now the program does not even exist). This is certainly more interesting than useful, but it might be worthwhile to conduct further research on how NFL teams overrate or underrate prospects out of small schools in the NFL draft.

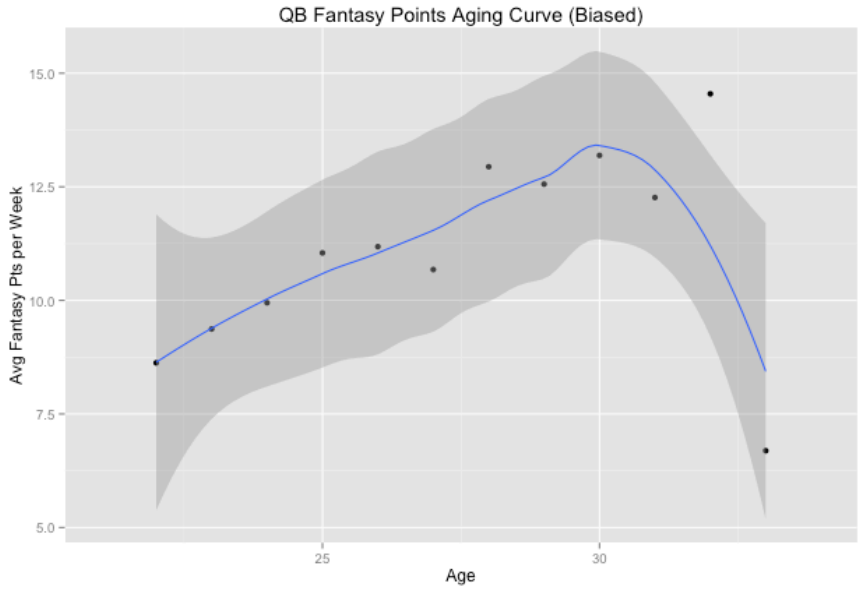
There is also a very weak positive correlation between age and weekly fantasy points for QBs, a weaker positive correlation between age and weekly fantasy points for WRs, a still weaker positive correlation between age and weekly fantasy points for TEs, and basically no correlation between age and weekly fantasy points for RBs. Much of the positive correlation for each position is likely due to survivor bias - those players who are still in the league at a later age are probably better than those players who are no longer in the league at that age.

There is a much stronger correlation between experience and weekly fantasy points for each position than there is between age and weekly fantasy points for each position. This follows the reference to survivor bias, as those players who are still in the league for many seasons are probably better than those players who do not last many seasons in the league.

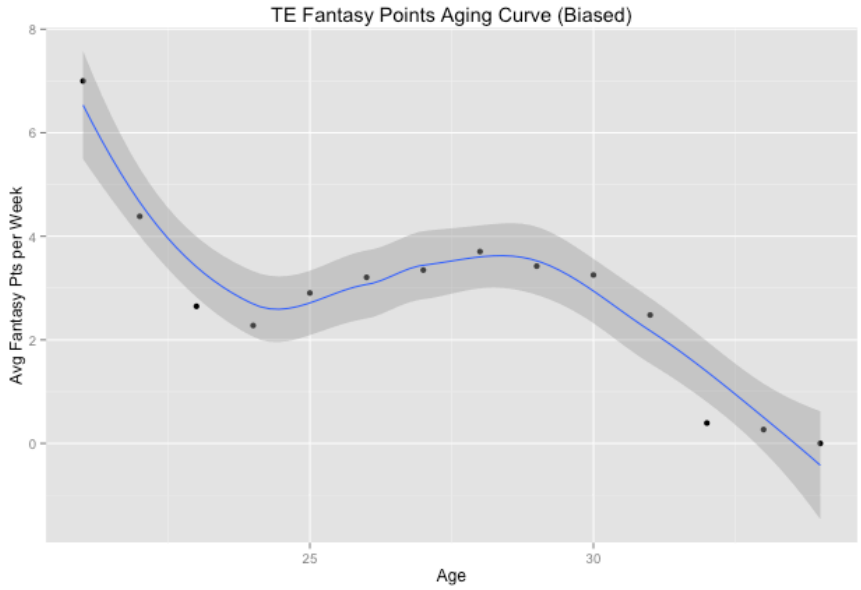
We can visualize this effect by graphing the “aging curve” for each position (Figures 23, 24, 25, 26). The average weekly fantasy points falls on the y-axis and age (binned to the official league year age) during a given season is on the x-axis. While this is helpful for visualizing how players at different positions improve, peak and decline in terms of their fantasy performance,

it is very important to be wary of entry bias and survivor bias (the best players play early in their career and the best players play to a later age). In particular, there were very few TEs who played in the league before age 23 in the timeframe of the dataset. Therefore, in the TE aging curve graph (Figure 24), the values at age 21 and 22 are unduly influenced by Aaron Hernandez, who was one of the best fantasy TEs in the game before being arrested.

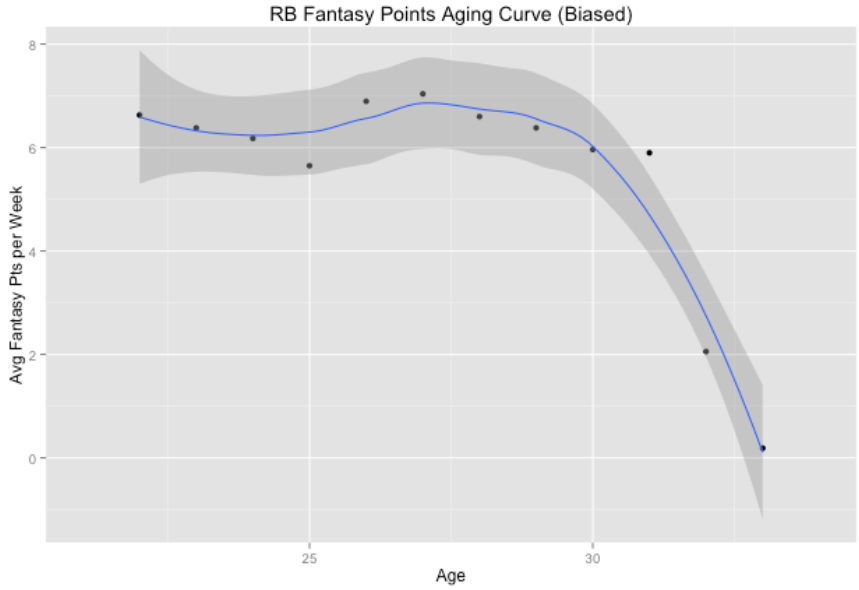
[Figure 23]



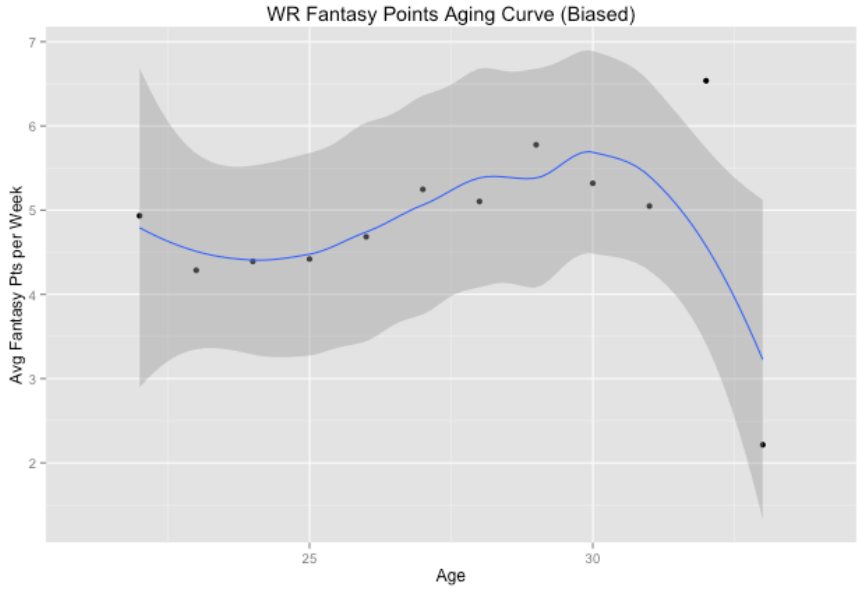
[Figure 24]



[Figure 25]



[Figure 26]



Despite these limitations (which are inherent in any aging curve analysis), there are still some very interesting and practical takeaways for anyone who wants to better understand fantasy football and employ a better strategy in their leagues. We can clearly see improvement, peak and decline for QBs, WRs and RBs (and TEs by looking from age 23 onward). And most interestingly, decline for RBs is much sharper than for other positions. RB fantasy production has tended to fall off of a cliff as the player in question approaches 30 years old. This intuitively makes sense as RBs take the most beating on their bodies and have a high risk of injury, which at older ages can even end the player's career. Anyone who drafted RB Frank Gore prior to the 2014 fantasy football season will agree. Gore was one of the most productive fantasy football RBs during the last few seasons, but at age 31, he has been a big disappointment for his fantasy

owners this season as he has averaged only 86 fantasy points as of Week 13, fewer than 24 other RBs.

We can tie these analyses together with a regression approach to predict weekly WR fantasy points from a player's draft position, age and age when drafted. In our exploratory analysis, we saw quite clearly that the distribution of fantasy points, even when split by position, is very skewed to the right (Figure 15). Therefore, it is important not to take the results of a linear regression model literally (Figures 27 and 29). Nevertheless, the results are useful for constructing one's fantasy football team.

The skewness of the WR data also lends itself well to a Poisson regression model (Figures 28 and 29). To account for the Poisson model assumptions, all fantasy performances of fewer than zero points were coerced to zero.

[Figure 27]

```
WRlm1 <- lm(fpts ~ Pick + age + Age_When_Drafted, data = WRtrain)

summary(WRlm1)
# Call:
#   lm(formula = fpts ~ Pick + age + Age_When_Drafted, data = WRtrain)
#
# Residuals:
#   Min     1Q Median     3Q    Max
# -7.798 -3.863 -1.855  2.352 36.071
#
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
# (Intercept)      10.969809   1.655394   6.627 3.65e-11 ***
#   Pick            -0.018422   0.000972 -18.952  < 2e-16 ***
#   age              0.266160   0.031440   8.466  < 2e-16 ***
#   Age_When_Drafted -0.505412   0.080642  -6.267 3.86e-10 ***
#   ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 5.629 on 8003 degrees of freedom
# Multiple R-squared:  0.06552,  Adjusted R-squared:  0.06517
# F-statistic:   187 on 3 and 8003 DF,  p-value: < 2.2e-16

# the R-squared is very low, as expected, because the dependent variable is
# weekly fantasy points, rather than full-season fantasy points

# compute the in-sample RMSE of the model
sqrt(mean((WRlm1$fitted.values - WRtrain$adj_fpts)^2))
# 5.626569
```

[Figure 28]

```
WRpois1 <- glm(adj_fpts ~ Pick + age + Age_When_Drafted, data = WRtrain, family = poisson)

summary(WRpois1)

# Call:
#   glm(formula = adj_fpts ~ Pick + age + Age_When_Drafted, family = poisson,
#       data = WRtrain)
#
# Deviance Residuals:
#   Min       1Q   Median       3Q      Max
# -4.1580  -2.4211  -1.0860   0.9989  10.4908
#
# Coefficients:
#   Estimate Std. Error z value Pr(>|z|)
# (Intercept)       2.736e+00  1.364e-01   20.06   <2e-16 ***
#   Pick             -4.413e-03  9.047e-05  -48.77   <2e-16 ***
#   age               4.928e-02  2.429e-03   20.29   <2e-16 ***
#   Age_When_Drafted -9.292e-02  6.589e-03  -14.10   <2e-16 ***
#   ---
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# (Dispersion parameter for poisson family taken to be 1)
#
# Null deviance: 52345  on 8006  degrees of freedom
# Residual deviance: 48423  on 8003  degrees of freedom
# AIC: 67689
#
# Number of Fisher Scoring iterations: 6

# compute the in-sample RMSE of the model
sqrt(mean((WRpois1$fitted.values - WRtrain$adj_fpts)^2))
# 5.612309
```

[Figure 29]

```
# compute the out of sample RMSE
WRpois1.predict <- exp(predict(WRpois1, newdata = WRtest))

sqrt(mean((WRpois1.predict - WRtest$adj_fpts)^2))
# 5.58728 # good - the model is not overfitting the training data

WRlm1.predict <- predict(WRlm1, newdata = WRtest)

sqrt(mean((WRlm1.predict - WRtest$adj_fpts)^2))
# 5.607048 # good - the model is not overfitting the training data
```

Through the resulting linear regression model as well as the resulting Poisson regression model, we can see that draft position is an extremely significant variable in terms of explaining WR fantasy production (Figure 24). While it is pretty obvious that WRs who get drafted earlier tend to be more productive fantasy players, this is nevertheless very important to keep in mind for anyone who wishes to optimize their fantasy football team.

We can also see that WRs who were drafted at younger ages tend to be more productive in fantasy. This could be for a number of reasons. All else equal, a younger player probably has a higher "true talent" level if his college production translates to an equivalent draft position compared to older WRs. Or it could be that WRs drafted at younger ages tend to be more naturally athletic, else their younger bodies would fail to show the requisite skill on the field equal to older WR prospects.

Finally, we can see that older WRs tend to be more productive in fantasy. This is no doubt in no small part because older WRs who are still in the NFL are probably much better than those WRs who retired before they reached an older age.

So how is all of this information and analysis useful for fantasy football players? We have shown that there are several components in analyzing player performance which fantasy team owners can use to improve their chances of success in their leagues: comparing positions, weather fluctuations, the draft (both real and fantasy), and age. These examples are only a few of the potential ways to get an advantage, but any analysis of fantasy football or projection of fantasy success must be taken with a grain of salt; fantasy football, like many games, includes a lot of randomness, and no one can predict future success or failure with total certainty. That randomness, along with new ways to get ahead, provides the enjoyment that keeps players coming back to the game each year, annually betting money even though they never seem to win.