Fantasy or Reality?

Investigating Popular Fantasy Football Draft Strategy

James Walsh

October 25, 2019

Springboard Data Science Career Track - Capstone 1 Report

"Fantasy football is about proving that you are better than your friends." - Rodney Ruxin

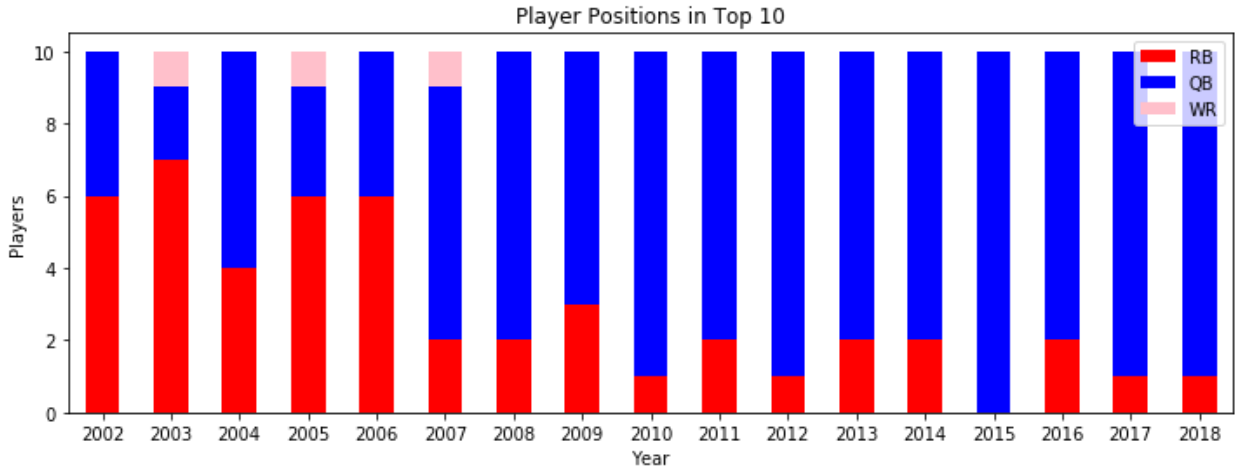
Anyone who plays fantasy football is familiar with two ubiquitous rules of thumb: 1) draft a running back first and 2) draft a kicker last. Even in 2019 when it is widely believed that the NFL is an increasingly pass-centric league, expert sources like ESPN continue to publish lofty projections where running backs are deemed top assets to be targeted early and often in fantasy drafts. However, a look at the top fantasy scoring players of 2018 suggests otherwise. Patrick Mahomes (quarterback) was the highest scoring player in 2018 with roughly 417 fantasy points. That is over 100 points more than the 2019 projected total for Saquon Barkley (running back) who is the consensus top fantasy draft pick for 2019. In fact, the aforementioned Mahomes is the highest projected quarterback for 2019 and sits at #43 overall on ESPN's fantasy draft projections, nestled between the #19 wide receiver and the # 21 running back. How can that be? Shouldn't the highest scoring players be targeted at the top of fantasy drafts? Using data acquired from fantasydata.com which collected player stats from 2002 through 2018, I set out to determine which position between quarterback and running back truly is the greatest asset to a winning fantasy team.

There is a lot of variation between fantasy leagues in terms of scoring and roster formats, so there will be some ambiguity to this question. So for the sake of simplicity, this project will factor in the following assumptions:

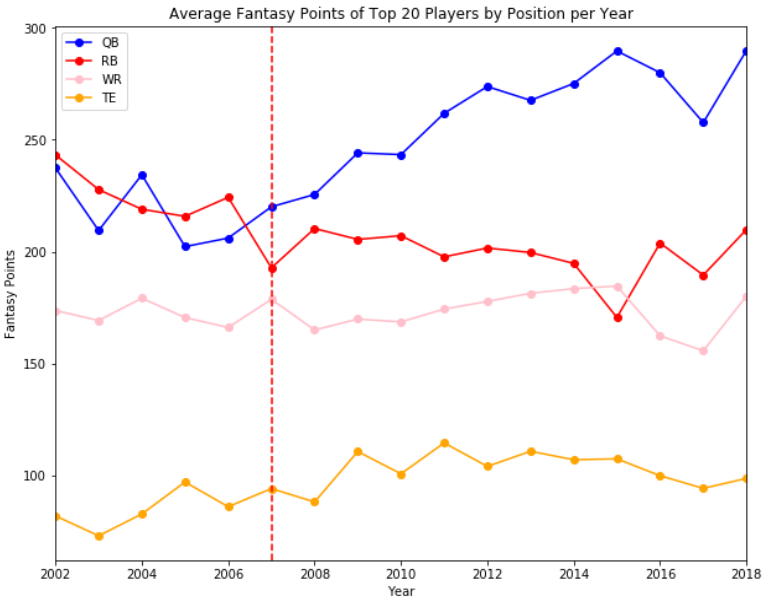
1. Standard scoring according to ESPN Fantasy. No points per reception, though that could affect the results of the analysis.
2. Fantasy team rosters allow for starting 1 quarterback, 2 running backs, 2 wide receivers, 1 tight end.
3. Real-world situations like player trades and injuries will impact total fantasy points but are not taken into consideration during analysis of the data.

The data gathered from **fantasydata.com** contains individual players by row and their accumulated real-world stats in the columns. It also contains their respective NFL team, position, average fantasy points per game and their total fantasy points for the season by which the players are ranked and ordered. I pulled the data from the website by year (2002-2018) and uploaded the csv's into my notebook where I also added a column for the year that data represents. After that, I merged all years into one pandas dataframe for easier accessing, manipulation and visualizing. From there, I began to inspect the data to make sure there were no considerable issues with the quality. One thing that jumped out immediately was that some players had an injury designation in their name column, such as "O" for "out", "IR" for "Injured Reserve" and "SUS" for suspended. I parsed over the column and removed all these designations leaving only the actual name of the player. Later, I noticed some players' positions were not correct e.g. Rich Gannon being labeled as a running back even though he played quarterback. To correct this, I searched for players whose stats did not match the position they played, like receivers passing for more yards than they caught and so on. Once these anomalies were identified, I updated their positions to be accurate. The data was pretty clean aside from these issues, so I began to explore and see what insights I could find with regards to my ultimate goal of determining whether quarterback or running back were really the most valuable position in fantasy football.

As previously mentioned, the highest performing fantasy player of 2018 was a quarterback. In fact, 9 out of the top 10 fantasy players of 2018 were quarterbacks. From 2002 through 2006, running backs were the far more prevalent position amongst the top 10. There were even a couple wide receivers here and there. However, 2007 marked a shift and the amount of running backs in the top 10 per season dropped substantially. Since that time, quarterbacks have dominated in terms of fantasy production. In fact, in

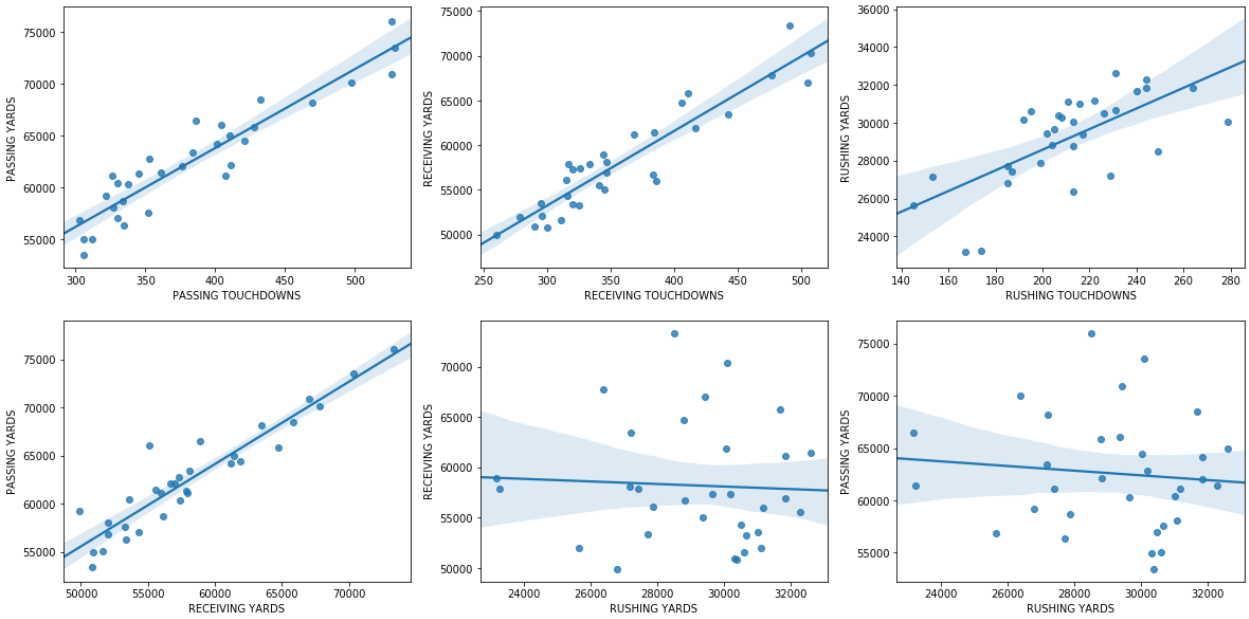


2015, the top 10 was *entirely* quarterbacks. Not only have quarterbacks been outperforming the other position groups (which begs the question; what happened in 2007 that caused this dramatic shift?), but their scoring has been continuously increasing over time. The average end-of-season fantasy point totals for quarterbacks have been rising in a manner that mirrors their dominance of the top 10 while running backs have seen a steady decline in fantasy scoring in recent years (again, 2007 stands out as a point of demarcation). Is there a statistical correlation? Performing a linear regression of running back against quarterback fantasy scoring averages yields a running back coefficient of -1.024 which indeed indicates a negative correlation. Meanwhile, wide receivers have remained relatively steady over this time and tight ends have seen only a slight rise in fantasy production, which is interesting considering the rise of quarterback production (more on that later). From this data, it is fairly obvious that quarterbacks have become the highest scoring individual players in fantasy football, but does that make them the most *valuable*?



Before investigating this question further, I explored what (if any) statistical patterns and correlations may exist within the data. Since this dataset is relatively small (~5000 rows, 17 columns), there are only so many features of the data we can compare with any statistical relevance. The correlation between player stats and their fantasy points is pretty obvious, so I explored the relationships between

the real-world stats that contributed to fantasy points. First, I grouped the data (passing yards, passing touchdowns, rushing yards, rushing touchdowns, receiving yards and receiving touchdowns) by their respective NFL team. Then, I used regression plots to visualize the trends. Some were as expected, like passing yards and passing touchdowns being closely correlated. However, the correlation between rushing yards and rushing touchdowns was somewhat low by comparison. This could indicate that a running back's best practical function is short yardage situations between the goallines with scoring coming frequently from those shorter runs and the infrequent long gain.

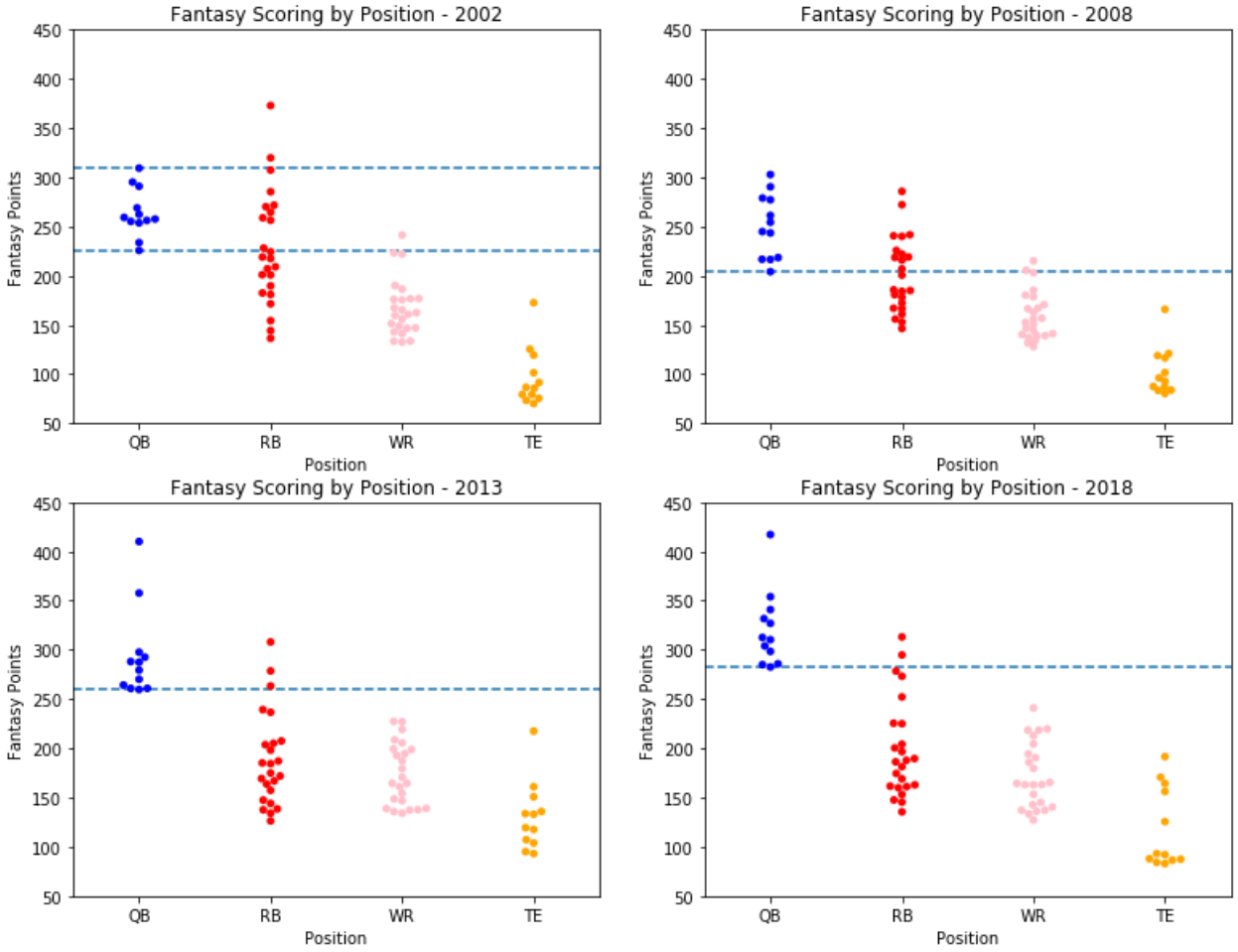


I also looked for other, less obvious correlations like that between passing and rushing and so on. As expected, passing yards and receiving yards were tightly correlated. However, there seemed to be no correlation at all between any of the other categories. It is commonly understood in the NFL that complementary rushing and passing games are critical to a team's on-field success. However, the fantasy data suggests that, from a statistical perspective, rushing seems to be completely independent from the passing game. While this does not address my question regarding whether quarterback or running back is the more valuable fantasy asset, it does offer some interesting insight into the data itself and suggests that their performances can be viewed as statistically independent of each other.

A key difference between the quarterback and running back positions in fantasy football is that, under standard formats, a team can start only one quarterback compared to at least two running backs (possibly more with flex spots). This means that in a standard twelve team league, there are twelve "starting" quarterbacks and twenty-four running backs. Therefore, the field for starting running backs is twice as wide and the variance of potential scoring within that group is higher as well. I created groups of the top 12 quarterbacks, 24 running backs, 24 wide receivers and 12 tight ends to note the differences amongst the scoring of these positions proportional to their usage in real fantasy leagues.

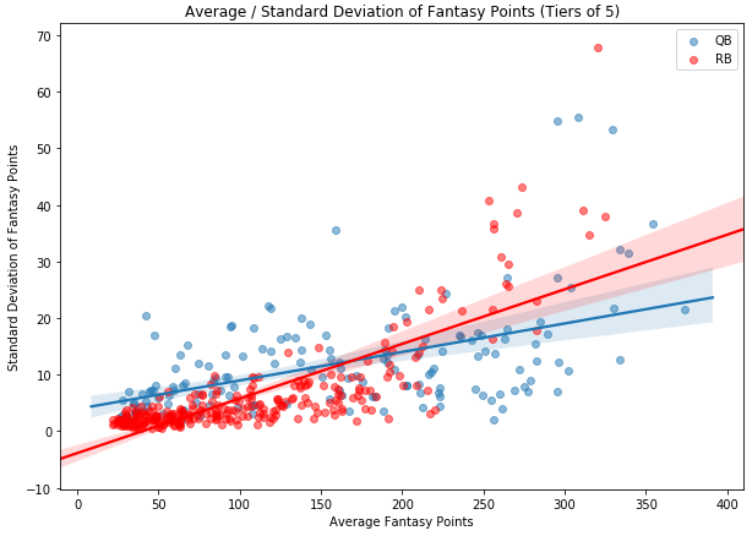
I plotted out the end-of-season fantasy point totals for players in these "starter" groups over four separate seasons (2002, 2008, 2013 and 2018) in order to capture any trends over that time. Overall, the wide

receiver and tight end groups' fantasy scoring has remained fairly consistent over the years, which is somewhat surprising given the rapid uptick in quarterback production. This could indicate greater quarterback totals are the result of spreading the passes around more than in previous years and that the primary beneficiaries of a pass-centric offense are not at the top of an NFL team's depth chart, but rather the second, third or fourth receiving options. So far as the quarterback / running back relationship goes, there is again a definite trend of the quarterbacks overtaking the running backs in terms of relative fantasy production. In 2002, the entire starting quarterback group was outperformed by two running backs. Also, there was a low variance amongst the group where running backs had a fairly wide distribution. By 2008,



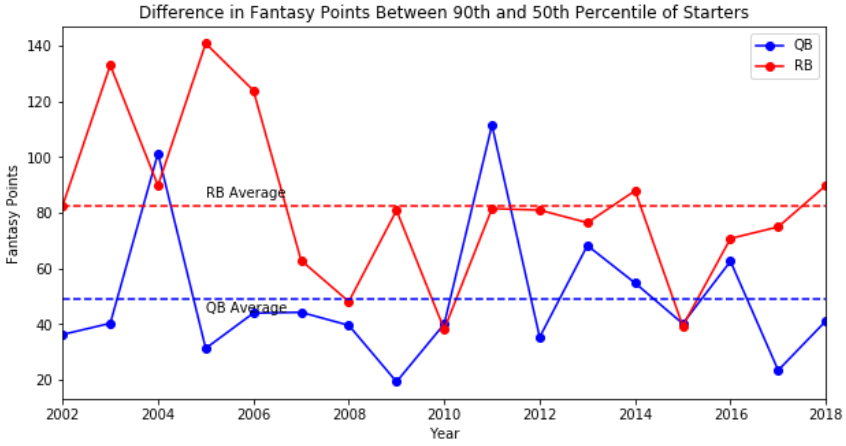
the top quarterbacks had begun to overtake the running back group in fantasy scoring and their distribution began to spread out more. The lowest performing starting quarterback ranked right around the middle of the pack for starting running back group whose distribution condensed considerably from 2002. 2013 depicts a continuance of this trend as only three running backs scored higher than the lowest scoring starting quarterback. One of those running backs outperformed ten of the twelve starting quarterbacks, but there were two rather extreme outliers amongst quarterbacks and the rest of the group regressed back into a tight distribution. By 2018, the running back group seemed to have changed little from 2013 in terms of overall performance and distribution, but the quarterback group as a whole seemed to spread out more where the vast majority of the starting running backs were scoring well below the lowest scoring starting quarterback.

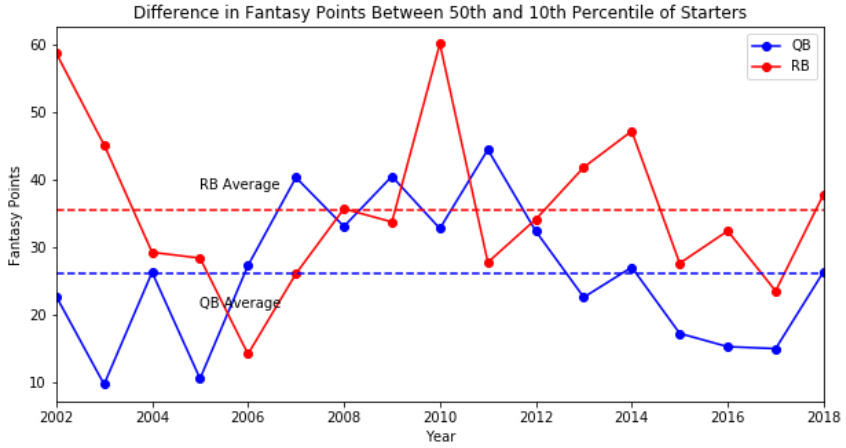
From this data, we can conclude that starting quarterbacks have, in fact, steadily overtaken starting running backs in terms of fantasy production which undoubtedly makes them a critical fantasy asset. However, we can also see that both starting groups appear to be rather bottom-heavy with only a small portion outperforming the rest. Are these outliers simply that; statistical anomalies that dilute the data? Since we're looking at relatively small data, I am inclined to believe these outliers are essential to understanding the true value of each position. As noted before, since the starting running back group is twice the size of starting quarterbacks, the variance amongst running backs is higher. To account for this, the question becomes which of these two groups not only scores more fantasy points, but scores those totals with greater consistency? To explore this, I placed each season's quarterbacks and running backs into groups of five and plotted those groups' total fantasy point means against their standard deviations. In the plot, greater along the x-axis indicates greater fantasy scoring means and lower along the y-axis means more consistency in that scoring. We can see a greater concentration of quarterbacks in the lower-right quadrant of the plot which indicates higher scoring with lower variance amongst top tier quarterbacks. Also, the regression line is less steep than that of the running backs which shows their score as a whole has lower variance. In fact, the running back tiers with the highest fantasy scoring are also those with the highest variance, while the majority of the running back tiers concentrate around low variance and relatively low scoring.



It is reasonable to infer that upper tier quarterbacks score more fantasy points with more consistency than their running back peers, but does that make really make them more *valuable*? After all, scarcity breeds value and if running backs that consistently perform at a high level are empirically more rare, would that suggest that top end running backs have more relative value than quarterbacks? In order to properly understand this value, we need to take into consideration how fantasy teams are commonly constructed: the draft. There are several popular draft strategies out there but as I mentioned in the introduction, it is a widely held notion amongst fantasy football participants and experts that running backs ought be targeted early and often in a fantasy draft. Let's say we have five quarterbacks with a mean fantasy point total of 350 and a standard deviation of 10. We also have five running backs with a mean fantasy point total of 300 and a standard deviation of 60. The upper limit of the running back position could be much more than 300 points and the lower limit could be much less, while all five quarterbacks' fantasy point totals would be much closer to the mean and thus each other. If you're picking first in this draft, should you go for the more rare but volatile upper tier running back or the more steady and consistent quarterback? Surely personal preference or strategy comes into play, but what does the data recommend?

To answer this, we need to understand how much variance there is within these two positions. Quarterbacks may net their team the higher individual fantasy point total, but are there more viable options within this position group than running backs? How much of a dropoff in production is there between upper tier players and the rest? We return to our "starters" (top 12 quarterbacks and top 24 running backs) and measure the difference between the 90th percentile and the 50th to give an indication of how the top performers at each position compared to the middle of their respective packs. With few exceptions, running backs experience a far greater dropoff in fantasy production than quarterbacks over the years. When comparing the 50th and 10th percentiles of these position groups, we see a similar trend with only a few cases of quarterbacks experiencing a greater dropoff in fantasy scoring than running backs. This shows that high performing running backs are indeed in short supply and while a quarterback may net a fantasy team more points than a single running back, the relative value that running back provides drops far more drastically than it does for quarterbacks as you progress down their respective ranks.



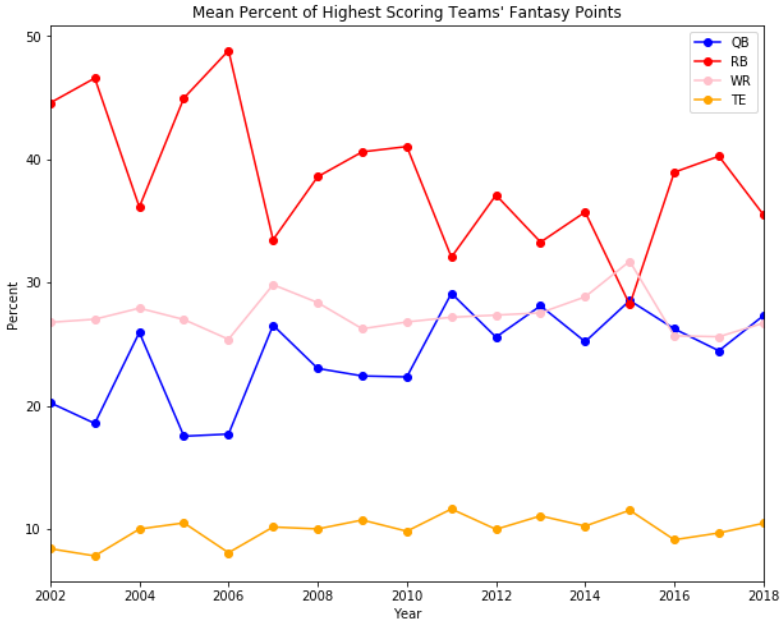


All of this provides valuable perspective into building a winning fantasy team. We've established that quarterbacks tend to provide more fantasy points on a consistent basis where the limited stock and volatile "boom or bust" potential of running backs leads experts to insist upon targeting them early in drafts. However, if you're playing fantasy football, your goal is to win your league. While the analysis thus far has focused on fantasy points and relative value between quarterback and running back, the answer regarding which of these two positions has the greatest influence on high scoring teams remains unclear. This is difficult to address scientifically since there is so much randomness and chance involved in fantasy football including drafting, weekly matchups, player injuries, player trades, etc. However, knowing the limitations of the dataset, I determined an interesting way of capturing the practical value of these positions was to simulate some form of 12 team league. Each team is comprised of 1 quarterback, 2 running backs, 2 wide receivers and 1 tight end with players selected from a particular year's end-of-season player rankings. I used a snake draft methodology wherein the drafting order is reversed

after each round until each team is full with the requisite positions. Since individual preference and strategy play a significant role in a fantasy draft and you never really know how they're going to pan out, I incorporated some variation into the process by making each pick a random selection within the top 9 players available and removing that player from the pool after they were picked. I ran this draft simulation 1000 times for each year from 2002 through 2018 and captured for analysis the team with the highest cumulative fantasy point total from each run. I ran one simulation with 2018 data that used no random choice and went strictly down the ranked order of players with the most fantasy points to see how a draft with perfect data would play out. Interestingly enough, the highest scoring team (if only by a fraction of a point) was that with the highest scoring running back of 2018, not the highest scoring quarterback.

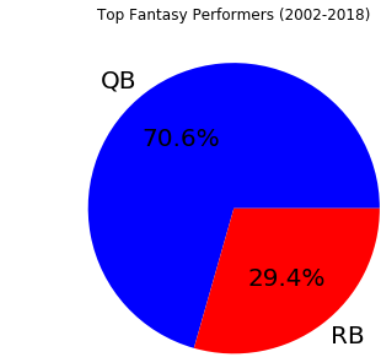
OVERALL RANK	NAME	POS	PTS	YEAR
6	Todd Gurley	RB	313.1	2018
14	Cam Newton	QB	282.6	2018
41	Travis Kelce	TE	191.6	2018
58	Stefon Diggs	WR	164.3	2018
69	Odell Beckham Jr	WR	153.3	2018
82	Matt Breida	RB	135.5	2018

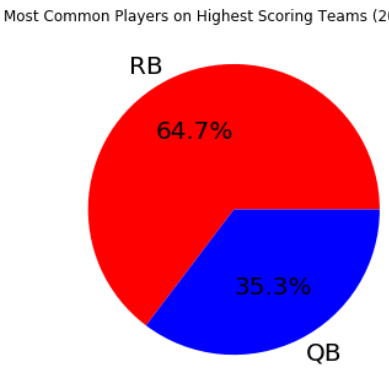
After I had collected the highest scoring teams out of each run of the draft simulation, I began my analysis on the composition of those teams in order to determine which position had the greatest impact towards making these teams the highest scoring in their particular "league." First, I compared the mean fantasy point totals for each position from the highest scoring teams for a given year. The trends here were very similar to those found when I calculated the average fantasy scoring of all four positions over the years, albeit with more extreme peaks and valleys. This shows that even amongst the highest scoring teams in my simulated leagues, quarterbacks outscored running backs, but aside from that it does not really illuminate anything I didn't already see in earlier analysis (see figure 2, page 2). Next, I took into consideration the fact that teams have 2 running backs compared to 1 quarterback and calculated the percent of the team's total fantasy points accounted for by that position as a whole. Understandably, this metric shows the two running backs on a team accounting for a greater percentage of the team's total score than its one quarterback. While this is not entirely unexpected, what is surprising is how the percentages of these two positions over the years seem completely inverse of each other. Since we are looking at percentages, it makes sense that some positions rise where others fall, but it is remarkable the degree to which quarterback and running back seem to influence each other where wide receivers and tight end remain very steady over this time.



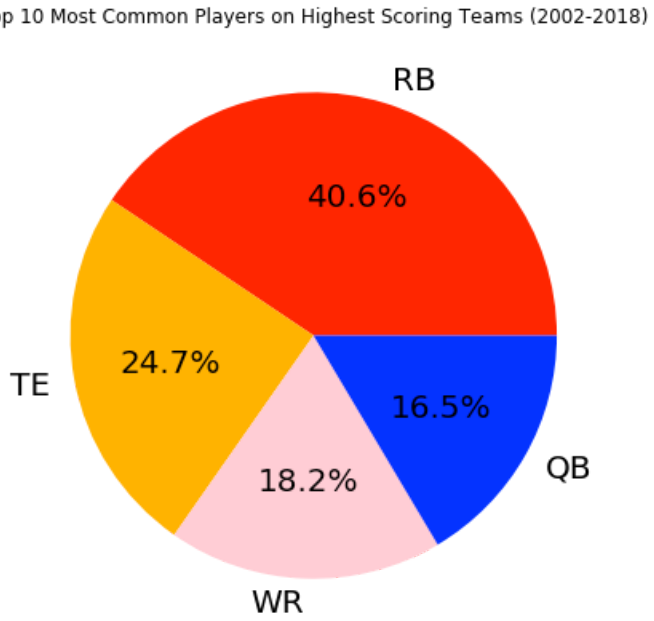
Aside from this intriguing correlation, the fact that the running back position as a whole accounts for a greater percentage of the highest scoring fantasy teams' final scores than the quarterback position is easily explained by the 2:1 ratio of running back to quarterback on a team. Also, since the mean fantasy points scored by each position group on the highest scoring teams echoed the mean scoring of those positions as a whole over time, we still lack a clear indication of which position has the greater influence on the highest scoring teams. In order to quantify this, I determined the individual players (and their positions) that were most frequently found on the 1000 highest scoring teams in each of the 17 seasons I simulated. Naturally, the ultra-high performing outliers in a season would likely be the most prevalent players on the highest scoring teams, but determining whether quarterback or running back is more likely to be the outlier that vastly outperforms everyone else is a key factor in that position's value and impact. For example: LaDainian Tomlinson (running back) was the top performer in the 2006 season when he netted the most fantasy points of any single player over the last 17 years. For context, he scored 28 rushing touchdowns (still an NFL record), rushed for over 1800 yards, caught passes for over 500 yards and 3 touchdowns and even passed for 2 touchdowns. His total fantasy points that year (425.1) were nearly 100 more than the #2 player. In my 1000 simulated leagues for 2006, Tomlinson was present on 82.4% of the highest scoring teams. The next most common player was Antonio Gates (tight end and Tomlinson's teammate) who was on 11.6% of the highest scoring teams in my 2006 simulations.

This is one very extreme example, but it's not only running backs that enjoy these tremendous performances. The following 2007 season saw Tom Brady (quarterback) rise to the top of the fantasy rankings, outscoring the runner-up (Tomlinson) by nearly 90 fantasy points. Brady was on 50.0% of the highest scoring teams from my simulations of 2007 where Tomlinson was on 25.8%. Despite quarterbacks overwhelmingly outscoring running backs in recent years (a quarterback has been the highest scoring fantasy player in 12 of the last 17 seasons i.e. 70.6%) seeing what players are the most frequent amongst the highest scoring teams from my simulation gives the clearest picture of what position gives their fantasy team the best chances of claiming that top spot. Through the results of my simulations, a running back was the most common player on the highest scoring teams in 11 (64.7%) of the last 17 seasons. In the remaining 6 (35.3%), a quarterback was the most common player. Despite quarterbacks frequently being the highest scoring players over the last 17 seasons, it is the top running backs that are most common amongst the highest scoring teams. A great example of this dichotomy can be seen in the 2009 data where Aaron Rodger (quarterback) and Chris Johnson (running back) were essentially tied in fantasy scoring with 342.96 and 342.90 points respectively. Both were nearly 50 points higher than the next player. However, Johnson was on 60.4% of the highest scoring simulated teams and Rodgers was on 21.4%. This trend can be seen as recently as 2017 where Russell Wilson (quarterback) ranked #1 in fantasy points (347.9) and Todd Gurley (running

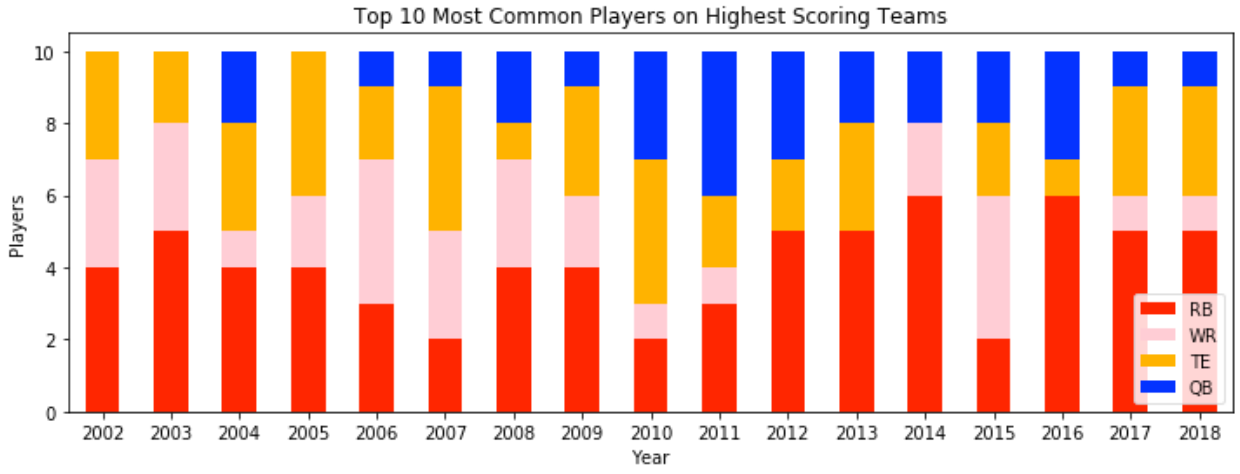




back) ranked #2 (319.30), but Gurley was on 57.2% of the highest scoring teams that year where Wilson was on 19.9%. This shows that the highest scoring player in a given season is not necessarily going to be the most common player on the highest scoring teams according to my simulation. In fact, if we look at the top 10 most common players on the highest scoring teams from each year I simulated, the data continues to favor the running back position which accounted for 40.6% of those spots. Interestingly, tight end was the second most common position in this grouping at 24.7%. While we have not focused on tight end much in this analysis, we have seen how this position is one that is frequently bottom-heavy where only a very small proportion of the players drastically outperform the rest (see page 4.) Much like running back, it is likely the relative scarcity of high performers at this position that makes them common amongst the top teams. Quarterback was the least frequent position in this grouping at only 16.5%.



Analysis of the results from my draft simulations certainly seems to indicate that running backs are the most common players amongst the highest scoring teams in a season. This is a clear indication that running back serves as the greatest asset for building a top scoring fantasy team. Even in 2018 where the NFL is viewed as a pass-centric game and quarterbacks are receiving record contracts, 5 of the top 10 most common players amongst highest scoring fantasy teams were running backs, 3 were tight ends and only 1 (albeit *the* most common) was a quarterback. In fact, despite quarterbacks overtaking running backs as the highest scoring players (see figure 1, page 2), running backs have steadily risen as the most common players amongst the highest scoring teams. Again, this is likely due to the relative scarcity of high volume, "workhorse" running backs in an era of quarterback driven, pass-centric NFL offenses.



The simulations used here are not perfect representations of actual fantasy football leagues and having the highest scoring team does not necessarily equate to winning in a standard format that uses win / loss records versus other teams in the league and a playoff structure. However, having the highest scoring team in your league will certainly improve your odds. To that end, the notion of running back being the top targets in fantasy drafts has been prevalent for many years and my analysis indicates that this holds true today. Even though quarterback fantasy scoring has risen in a manner that reflects their real-world proliferation over the years, the data indicates that this trend has actually capped their relative value from a fantasy perspective. As more quarterbacks in the NFL continue to put up ever-increasing on-field production, the supply of high performers at this position will continue to approach the demand of the fantasy leagues. Conversely, as the stock of running backs that are the focal point of their NFL offense has dwindled in recent years to the point that there are nowhere near enough to satisfy at least 24 fillable starting fantasy spots in a 12 team league, the value of this position remains as high as ever if not higher. As a side note, the analysis suggests that top tier tight ends have a sizable influence on high scoring fantasy teams for reasons of scarcity similar to running backs and are also worthy of early draft consideration. So, if you're picking early in your next fantasy football draft, confidently pick that running back and you'll be one step closer to a championship!

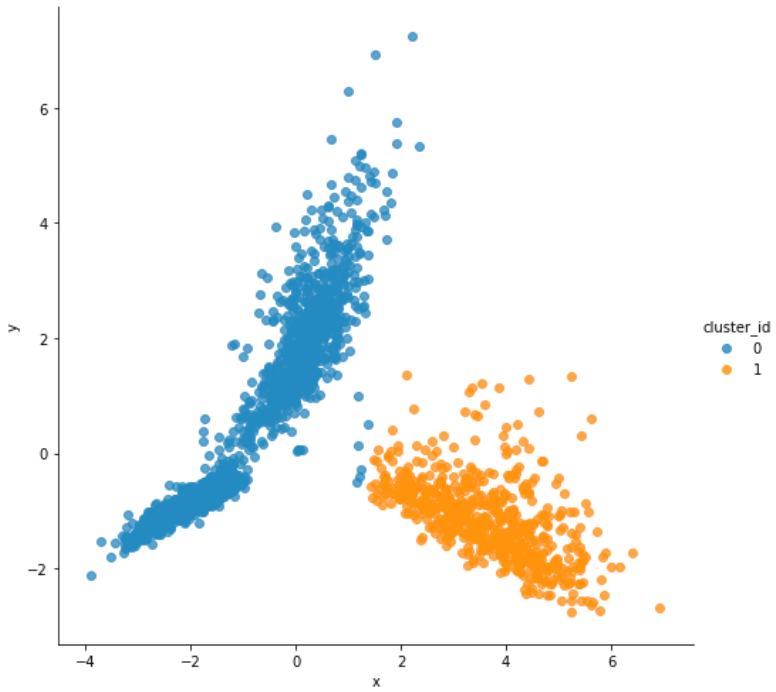
Now, should you really take a kicker last?

Clustering Analysis

With regard to my project's focus of determining which position in football, the quarterback or the running back, has the biggest influence on winning in fantasy football, the ML techniques to employ were limited. The data was very clean and fully labeled so there wasn't a lot of room for cleaning or improving data quality. There were also not a lot of features (17) to select for evaluation, so this was going to be a challenge.

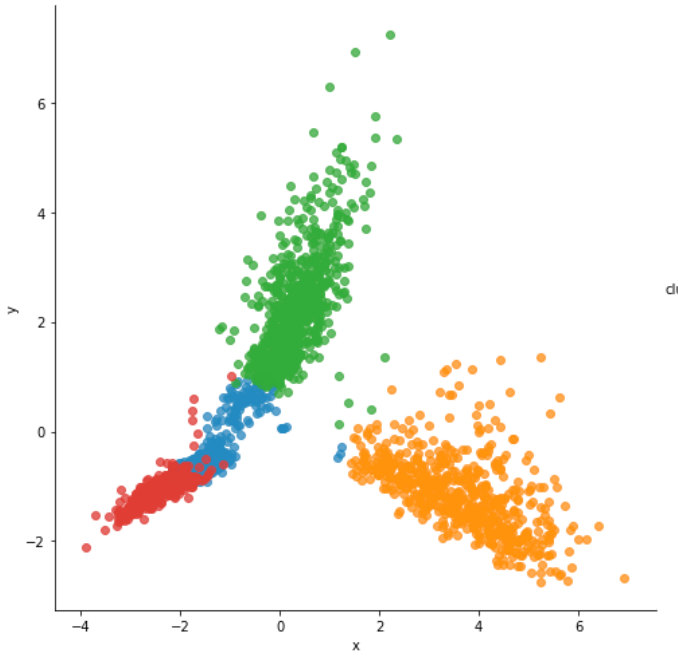
I took a shot with clustering to see if that would be able to identify and group players into their correct positions using k-means from scikit-learn. I removed all extraneous information from the dataset leaving behind only their names and stats for passing, rushing and receiving. Since there were 4 positions in the dataset I knew I wanted to see how successful k-means would be at generating 4 clusters, but I wanted to see what the optimal k-value for this data would be before attempting 4. First, I used the Elbow Sum-of-Squares method which indicated a k-value of 2 was best and given the size of the data, that did not seem odd. I also ran a silhouette test which identified 2 as the optimal k-value as well with a silhouette score of 0.746 by which "strong structure" is indicated. By contrast, a k-value of 4 yielded a silhouette score of 0.629 which indicates "reasonable structure."

Before I produced and visualized the clustering, I standardized the data with StandardScaler. Next, I used PCA for dimension reduction to enable 2D plotting. Then, I ran k-means using k=2 as suggested by the earlier Elbow and silhouette tests and produced the resulting plot to the right. The explained variance of this plot was 0.755, which is pretty good considering the breadth of the dataset. When I looked into the cluster groups, the algorithm seemed to label quarterbacks as 1 (orange) and the rest of the positions including some quarterbacks that had a decent amount of rushing yards as 0 (blue). In this case, it seemed to be passing performance that was the key feature of demarcation.



	0	1
PASS YDS	4.48	3136.65
PASS TDS	0.05	19.53
INT	0.02	11.67
RUSH YDS	339.08	148.95
RUSH TDS	2.42	1.43
REC	52.74	0.15
REC YDS	636.28	0.6
REC TDS	4.07	0.01
FUM LST	0.83	2.87
PTS / GM	9.19	14.37

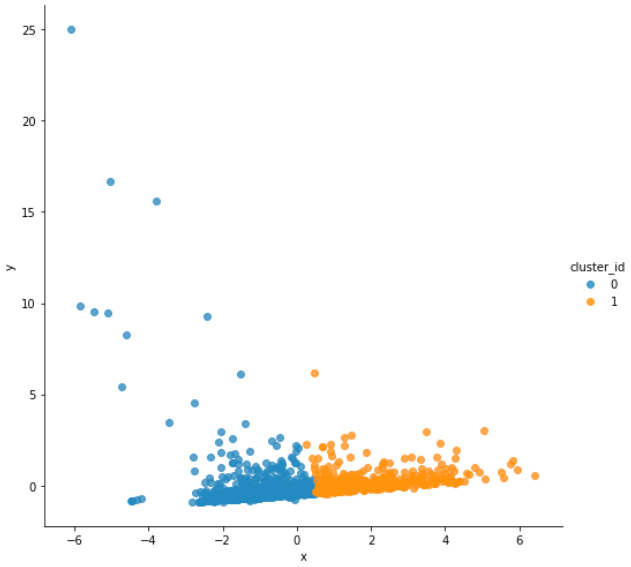
Next I ran a similar k-means using k=4 to see if it would split the positions appropriately. I ran StandardScaler and PCA as before but used k=4 and produced the plot to the right. Cluster 1 was relatively unchanged, losing only 2 players from the group. Here, 1 (orange) was again exclusively quarterbacks. 2 (green) was predominantly running backs with a few "mobile quarterbacks" that had a knack for running the ball, like Michael Vick and Tim Tebow. The difference between rushers and passers seemed reasonable, but I figured differentiating between receivers and tight ends with the data available would be tough since all it has to use is total receptions, yards and touchdowns. After analyzing the labels, it definitely seems that these were the parameters used to distinguish these clusters. 0 (blue) was composed receivers, tight ends and a few running backs that had notable receiving totals and



	0	1	2	3
PASS YDS	5.96	3143.41	8.51	0.91
PASS TDS	0.05	19.58	0.08	0.02
INT	0.02	11.7	0.03	0.02
RUSH YDS	56.32	147.72	912.08	20.51
RUSH TDS	0.27	1.43	6.71	0.09
REC	53.43	0.15	31.94	82.95
REC YDS	675.61	0.6	255.58	1137.85
REC TDS	4.2	0.01	0.98	8.53
FUM LST	0.47	2.87	1.47	0.58
PTS / GM	6.78	14.38	11.2	10.87

relatively low (>600) rushing yards. However, as a whole, this cluster had a mean receptions of 56.3, mean receiving yards of 675.6 and mean receiving touchdowns of 4.2. Group 3 (red) was composed receivers, tight ends, a few running backs that had more receiving yards than rushing yards and three quarterbacks that passed for under 1500 yards. This cluster had a mean receptions of 82.95, mean receiving yards of 1137.85 and mean receiving touchdowns of 8.5. Overall, it seemed Group 0 was the "good" receivers and tight ends with some competent pass-catching running backs and Group 3 was the more lackluster pass catchers along with comparatively high volume pass-catching running backs and a few uninspired quarterbacks. These two groups were understandably hard to recognize in cluster due to the limited data presented.

However, I still wanted to see if I could get better results, so I removed all quarterback and running backs and left only receivers and tight ends. Elbow and silhouette tests both indicated k=2 as the optimal value (perfect this time) so I ran k-means on this data to see what we got. This time, the results seemed to follow the previous model where "good" performers were in Group 1 (orange) and mediocre to poor



	0	1
PASS YDS	1.4	0.97
PASS TDS	0.02	0.02
INT	0.01	0.02
RUSH YDS	19.5	15.74
RUSH TDS	0.12	0.06
REC	54.84	84.65
REC YDS	716.12	1168.19
REC TDS	4.63	8.54
FUM LST	0.43	0.61
PTS / GM	6.86	11.01

performers were in Group 0 (blue). Also, the explained variance of this model was 0.489 which is not great.

I think the problem with these models is rather limited data. All I have to go on is passing, rushing and receiving stats which has been demonstrated to be sufficient and dividing quarterbacks, running backs and pass-catchers, but not great for understanding the difference between receivers and tight ends. For that to work better, I would need far more than the 10 features I have here and could include pertinent information like snap counts, formations used, targets, time on the field, etc.