The Search for Value:

A Statistical Analysis of Daily Fantasy Football Strategies

Dylan Farrell

A thesis presented for the degree of

B.A. in Statistics and Mathematics

Statistics Department

Harvard University

4/02/17

1 Abstract

This paper uses player data from the National Football League (NFL) and FanDuel, a daily fantasy sports company, in order to assess the strengths and weaknesses of common daily fantasy football lineup optimization strategies. While relatively new to the sports world, daily fantasy football has grown immensely over the last several years, with just under 4 million active users as of fall 2016.[And16] Several companies, most notably DraftKings and FanDuel, have sought to tap into this market and offer platforms that allow users to compete in tournaments and head-on-head match-ups against other users.[And16] In a nutshell, each DraftKings or FanDuel user competes by constructing a lineup of a set number of NFL players from different positional categories and entering that lineup into a weekly contest. Players in the lineup accrue points based on their statistical performances in the NFL games that week, and so each lineup ultimately has a total score representing the sum of all the fantasy points earned by the players in the lineup. While contest structures vary greatly, the common goal throughout each contest is to have a lineup that generates a higher total score than your competitors' lineups. Users cannot simply pick the best players in each position, however, as there are several constraints that each lineup must satisfy in order to be valid. The primary constraint is that the total salary of all the players in a lineup must fall under a specified salary cap. These are not the players' actual NFL salaries, but rather artificial fantasy salaries that platforms such as FanDuel generate for every player each week of the season. The formulas for generating these salaries are often complex, but in general higher quality players will tend to have higher salaries relative to lower quality players.[Zel]

While some users have their own models and lineup building processes, many users construct their lineups based on feel and on some common principles. One common lineup optimization approach is to attempt to maximize "Value", which is defined as $\frac{\text{Proejected Points}}{\text{Salary}}$.[Spr15] Thus this strategy essentially maximizes a lineup's estimated "bang for buck." There are many other strategies, including choosing players who are thought of as having a "high floor" or a "high ceiling."[Spr15] A player with a "high floor" is one who has a very low chance of putting up a sub-par performance that week, even if there's a fairly low chance they have a great performance. Meanwhile a player with a "high ceiling" is one who has the potential to put up a lot of points in the given week, but also might have a decent chance of having a sub-par score.

However, while terms such as these might appear in write-ups about specific players for a given week, they do not appear in statistical lineup optimization strategies. Most of the lineup optimization approaches focus purely specific point projections for each player's performance in a given week, and almost none of them consider a player's estimated points distribution for a given week

or try to quantify what exactly a given player's floor or ceiling is. This thesis fills the gap in the existing daily fantasy football lineup optimization strategies literature by creating estimated points distributions for every available player in each week of the NFL season between 2011 and 2017 and optimizing lineups based on specific characteristics of those distributions, such as the mean, median, variance, floor, ceiling, and others.

Research yields the following results for lineups consisting of one quarterback, two running backs, and three wide receivers with a salary cap of $40,000:

1. Optimizing for any of the mean, median, variance, floor, or ceiling of a distribution yielded a higher lineup score on average than a model than randomly picks players whose team salary is within $2,000 of the salary cap.
2. Optimizing for the mean or median of a distribution yielded the best results on average
3. Optimizing for the ceiling or variance of a distribution yielded the most results in the top 10th percentile of lineup scores.
4. Optimizing for the floor of a distribution yielded the least amount of results in the bottom 10th of lineup scores, but had very few scores in the higher percentiles.

2 Acknowledgements

Special thanks to Professor Parzen for agreeing to be my thesis advisor and for assisting me in developing methods for estimating player performance and comparing optimization strategies. Thanks also to Daniel Alpert for his insightful feedback and commentary throughout all stages of the production of this thesis.

Contents

3 Introduction to the NFL and Daily Fantasy Football

From the preseason games in August to the Super Bowl in February, each season the National Football League (NFL) is a source of fascination for fans and players alike. While many readers likely have a acquired at least a rudimentary understanding of the workings of the league, I take the following sections to briefly explain the basics of the NFL structure and game play.

3.1 NFL League Structure

The NFL consists of thirty-two teams split across two conferences and eight divisions. Each team plays sixteen games over a seventeen week regular season that spans from early September to the end of December. At the end of the regular season, the four division winners plus two additional wild-card teams from each conference make it to the playoffs and compete in a single elimination tournament over the span of five weeks, culminating in the Super Bowl each February.

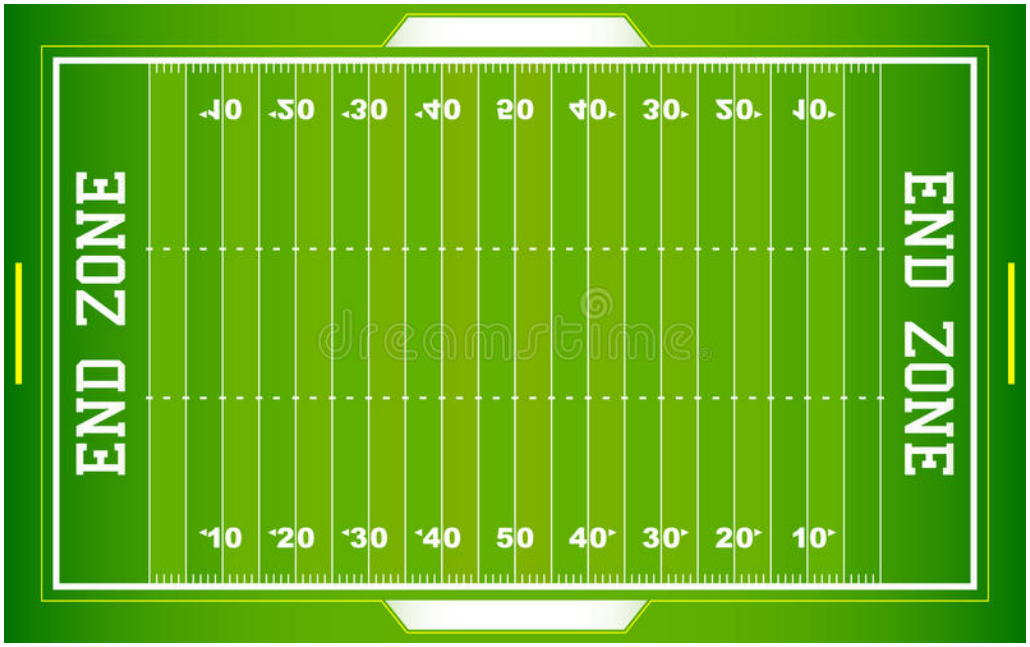
Figure 1: NFL League Structure: Conferences and Divisions

3.2 NFL Game Rules

In each NFL game, the two competing teams (each with eleven players on the field at a time) each try to move the ball down the field and score points. Each game consists of four fifteen minute quarters with an additional ten minute overtime if the teams are tied at the end of regulation. Teams can score in several ways. If a team crosses the plane of the opposing team's end zone while in possession of the ball, they score a *touchdown*. This usually occurs on offense, but teams can also score defensive touchdowns if the other team turns the ball over and the defense returns it into the other team's end zone. After scoring a touchdown, teams have two options: taking a *point after attempt* (also referred to as an *extra point*) and trying to kick the ball through the uprights

at the back of the end zone; or attempting a *two point conversion* and getting one try to cross the end-zone of the other team again, starting at the five-yard line. While on offense, teams can also score points by kicking *field goals*, in which a player on the offensive team attempts to kick the ball through the uprights at the back of the opposing team's end zone. The only other way a team can score points is when their defense tackles the opponent inside the other team's own end zone in what is called a *safety*.

Figure 2: Aerial View of a Football Field



A breakdown of the scoring plays and the number of points awarded for each is listed below:

Play Type	Number of Points
Touchdown	6
Point After Attempt	1
Two-Point Conversion	2
Field Goal	3
Safety	2

3.3 Football Positions

Each NFL team has a 53-man roster during the regular season and a 46-man game-day roster for each individual game. These players can play in a variety of positions on each of the three sides of the ball: offense, defense, and special teams. The offense attempts to move the ball down the field and score touchdowns and field goals. The defense tries to prevent the other team's offense from scoring. Meanwhile special teams players are involved in any plays with field goals, kicks, and punts.

While there are many positions in football, the relevant ones for daily fantasy football are: quarterback, running back, wide receiver, tight end, kicker, and defense as a whole. With the exception of defense and kicker, these positions are commonly referred to as *skill positions*, as they are the players who are actually in possession of the ball at any given time.

Quarterbacks (QBs) are the primary passers for their teams and attempt to throw the ball down-field to wide-receivers, tight-ends, and pass-catching running backs. They also can hand the ball off to running backs or keep it themselves and attempt to scramble down-field for yards.

Running backs (RBs) come in several different types. Power backs run in the middle of the field and use their strength to bulldoze their way for yardage.[Sil17] Speed backs are quick on their feet and use their speed to get to the edges of defenses and make big rushes for a lot of yards.[Sil17] All-purpose backs are good at running, catching, and blocking and can play a significant role in the running game and the passing game.[Sil17]

Wide receivers (WRs) attempt to create separation from defenders so that they can catch the ball down-field and then run for additional yards after the catch.

Tight ends (TEs) are involved mostly in blocking but can also get receptions and score touchdowns in the passing game.

Kickers (PKs) attempt to kick the ball through the uprights in field goal attempts and point after attempts and also kick the ball to the opposing team during kickoffs.

Defense (DEFs) is the set of players (eleven on the field at once) who attempt to prevent the opposing team's offense from scoring.

3.4 Team and Player Statistics

Over the course of each NFL game as teams score points and amass passing yards, rushing yards, and a variety of other statistics, players also accumulate individual statistics. The table below outlines the key statistics in each phase of the game:

Phase	Relevant Statistics
Passing (Primarily QBs)	Passing Attempts, Yards, Touchdowns, Interceptions
Running (Primarily RBS)	Rushing Attempts, Yards, Touchdowns
Receiving (Primarily WRs and TEs)	Receiving Targets, Receptions, Yards, Touchdowns
Kicking (Primarily PKs)	Field Goals, Extra Points
Defense	Points Against, Passing Yards Against, Rushing Yards Against

3.5 Fantasy Football

While NFL coaches and players are only focused on winning games and competing for Super Bowls, fantasy football players take a keen interested on these in-game statistics for individual players and defenses.

The first rules for fantasy football were initially postulated in March 1962 by Wilfred Winkenbach (a limited partner in the Raiders NFL organization), Bill Tunnell (the Raiders PR-man), and Scotty Stirling(a reporter for the Oakland Tribune) on a Raiders east coast trip.[Bro16]

The general idea was as follows: ordinary fans could join "fantasy leagues" in which they drafted teams of a set number of current NFL players from specific football positions. Each NFL player could only be on one team's roster. League members would match up each week of the NFL season and would gain fantasy points as players on their teams accumulated statistics in the actual NFL games. While rules would vary according to the league, players would earn a certain number of points for each passing yard, rushing yard, reception, touchdown, etc. Whoever had the lineup that produced the largest amount of fantasy points would win that week's match-up. And so over the course of the season, league members would develop win-loss records, and the member with the best record at the end of the season would be crowned the league champion.[Bro16]

From these humble beginnings, fantasy football soon blossomed into a nation-wide phenomenon, with over one million players in the U.S. by September 1989.[Bro16] Over the next decade it continued to grow in popularity, and in June 1999, Yahoo became the first major company to offer fantasy football for free as part of its site.[Bro16] This expanded the growth of the industry even further, resulting in over 12 million fantasy football players in the U.S. by September 2006.[Bro16]

3.5.1 Daily Fantasy Football

While normal fantasy football as described above is still immensely popular, a new form of the game called "daily fantasy football" came onto the scene in the early 2000s and has seen enormous growth ever since. There are several "platforms" for daily fantasy football, but the two most popular today are FanDuel (created in 2009) and DraftKings (created in 2012).[And16] Both of these companies are daily fantasy sports contest providers and run websites that allow users to participate in daily fantasy football each week of the NFL season.

So what exactly is daily fantasy football? In daily fantasy football, users assemble new lineups every single week of the NFL season as opposed to just once at the start of the season. Each week, these users can pay entry fees to enter these lineups into a variety of contest types. While contest types vary, in general lineups that perform well will earn money, while those that do poorly will lose money. Every user has the same players to choose from, and multiple users can roster the same player in a given week. Users cannot simply pick the best players in each position, however, as there are several constraints that each lineup must satisfy in order to be valid. The primary constraint is that the total salary of all the players in a lineup must fall under a specified salary cap. These are not the players' actual NFL salaries, but rather artificial fantasy salaries that platforms such as FanDuel or DraftKings generate for every player each week of the season. The formulas for generating these salaries are often complex, but in general higher quality players will tend to have higher salaries relative to lower quality players. The challenge for daily fantasy football players is figuring out how to create the lineup that they believe will score the most points while still staying under the salary cap.

3.6 Fantasy Lineup Construction

While NFL game-day rosters consist of 46 players, daily fantasy football lineups typically consist of 9-10 players.[rul] For example, FanDuel teams consist of the following positions: one quarterback, two running backs, three wide receivers, one tight end, one kicker, and one defense.[fan] The players do not have to be on the same team, and in fact in order to reduce correlation between players, FanDuel limits users to having a maximum of four players from the same team in their lineup.[fan]

Below is an example fantasy lineup for week one of the 2017 season:

Position	Player	Salary
Quarterback	Tom Brady	$8,900
Running Back	Kareem Hunt	$6,400
Running Back	Mark Ingram	$6,400
Wide Receiver	Danny Amendola	$5,700
Wide Receiver	Nelson Agholor	$4,900
Wide Receiver	Keenan Allen	$7,000
Tight End	Charles Clay	$4,600
Kicker	Graham Gano	$4,700
Defense	Jaguars	$4,100

The players in this lineup come from a number of teams and have a total team salary of $52,700, well under the FanDuel salary cap of $60,000.

3.7 Daily Fantasy Salary System

Daily fantasy football platforms such as DraftKings and FanDuel have complicated algorithms that assign a salary for every available player each week. According to Alex Zelvin, a part time FanDuel employee, the primary concern for most sites when generating salaries is to "Make sure that prices are set in a way that will make enough players equally appealing options that lineup duplication is minimal."[Zel] Each site picks a set salary cap and then generates salaries for each player such that there are many lineups that are close to hitting the cap that are equally desirable.

This is an interesting approach, because while it makes sense that daily fantasy platforms want to ensure that users have as diversified a set of lineups as possible, the fact that the salaries are set to factor in user sentiment in addition to player matchups and historical performance means that there is an opportunity to find under-valued players who will predictably overproduce relative to their salary level.

3.8 Contests

FanDuel has several different types of contests that users can enter lineups into. Three of the most popular ones are head-to-head, 50/50s, and tournaments.

In head-to-head matchups, users pick a single other user who they would like to compete against. The user with the lineup that performs better that week wins a percentage of the other player's

entry fee.[dfab] Strategies for head-to-head are rather complex, since the optimal strategy is conditional on what your opponent's strategy is. For example, you might assume that your opponent is going to fill his/her roster with high variance players who could put up a lot of points but also might score very few points. In this case, you would probably want to take a more conservative approach and build a lineup of players with high floors so that you can expect your lineup to produce at least a moderate amount of fantasy points. However, if your opponent is picking a lot of players with high floors...

A second popular contest format is 50/50 matchups, in which multiple users enter lineups and the top 50% of lineups each win the same amount of cash while the bottom 50% all lose their entry fees.[dfaa] This type of match-up is enticing for many users, especially inexperienced ones, because they only have to beat the average entry and thus are not as worried about having to compete against professional players as they would be in a head-to-head match-up.

A third popular format are tournaments. In tournaments, multiple users enter lineups and the best lineups win cash in a very top heavy structure. Tournaments are exciting for many players because they offer the potential to win large amounts of cash.[dfac] The first place user in some tournaments can win up to one million dollars, while the rest of the top ten all win thousands to tens of thousands.[dfac] However, given the large number of entries in these tournaments, it is very very difficult to place near the top. To be a top finisher in these tournaments, you have to be enough of a contrarian to create a unique lineup while also finding a very high performing one, a task much easier said than done.

4 Lineup Optimization Background and Literature Review

4.1 Player Terminology

As daily fantasy sports have evolved a number of terms have come into use to describe players and their potential in a given week. A number of desirable traits for players in a lineup are:

Upside: players with upside are those that have low salaries but have potential to put up a lot of points.[Spr15] For example a second string running back will likely have a low salary, but is also one injury away from being in a starting role. Thus this player has upside, especially if you have good information that there is a decent chance of the player getting more snaps than usual in the coming game. For example, in 2017, Alfred Morris was a moderately talented running back on

the Dallas Cowboys, but he was buried on the depth chart behind Ezekiel Elliott, one of the best running backs in the game. However, throughout the course of the season, Elliott was in the midst of arbitration for his appeal of a suspension that had been handed down to him by the league. While the start of the suspension kept being delayed by the court of appeals, it seemed clear that at some point in the season Elliott would likely be forced to miss some games. This made Morris a player with a lot of upside, especially in week 11 when Elliott was denied his appeal and was forced to serve the first game of the suspension. Morris had 27 attempts for 127 rushing yards and one rushing touchdown that week, which was a performance well above his year-to-date average in all categories.[pro]

Consistency: as the name suggests, a player with consistency is one who puts up a similar number of fantasy points each week and can be reliably expected to produce in the same small range every week.[Spr15] Of course, consistency in itself is not necessarily beneficial for a lineup since many players consistently put up very few points. For example, Russell Wilson was one of the most consistent players in the NFL last season who averaged 21.7 fantasy points per game with a standard deviation of 7.5 points.[pro]

Floor: a player with a high floor is someone who might not put up the largest number of fantasy points, but can at the very least be expected to put up above a certain amount even on a bad day.[Spr15] For example, in 2016, New England Patriots wide receiver Julian Edelman consistently caught six or more receptions and was a focal point for the offense, even while not catching as many touchdowns as other players on the team. Because he was such a key part of the Patriots offense, Edelman consistently put up over 6 fantasy points in that season, and thus was a player with a high floor.[pro]

Ceiling: a player with a high ceiling is one who has a realistic chance of scoring a large number of fantasy points in any given week, even if on average they only score a modest amount.[Spr15] For example, Kansas City Chiefs wide receiver Tyreke Hill has a remarkable amount of athleticism and has a decent chance of breaking free for either or a big reception or a big run for a touchdown at any point of the game. For this reason Hill has been one of the game's a higher-ceiling wide receivers for the past two years.[pro]

Value: players with high value are those who have a high expected number of fantasy points relative to their salary.[Spr15] Assuming the points predictions are accurate, the lineup with the highest total value will be the one that has the highest amount of points.

Of course a player that is consistent, has a high floor, and a high ceiling, and has a high expected number of fantasy points will likely have a very high salary as well, so it unlikely that that player will have great value.

Thus the art of daily fantasy football is determining which player traits are most important for the type of contest that the user is entering and then building a lineup optimized around selecting players with those traits.

For example, in 50/50 contests, all you have to do is beat half the field. Thus constructing a lineup filled with boom-or-bust players who have high ceilings but also have a decent chance of not putting up very many points is probably not a good strategy since there is no extra reward for having a high-scoring lineup once you're in the top 50%. A better strategy for this type of contest is to optimize for player traits that lead to lineups performing above average as much of the time as possible, even if this means that the lineup is never close to the top of the field.

Similarly, in tournaments, what really matters is not doing above average, but coming in as close to first as possible. For this reason, a good strategy is one that optimizes for whichever attributes will team that has the highest possible ceiling, even if the probability of the team hitting the ceiling is not that high.

4.2 Building an Optimized Lineup–A Modified Knapsack Problem

The most commonly employed numerical approach for constructing an optimized lineup is finding the lineup with the greatest value (expected points/salary) by using an optimization algorithm. Of course, as shown in this thesis, users can optimize for a "value" other than expected points/salary, such floor/salary or ceiling/salary.

Unfortunately, there are at least 30 available players in each position in daily fantasy football each week, so for a typical FanDuel lineup consisting of one quarterback, two running backs, three wide receivers, one tight end, one kicker, and one defense, there are over $\binom{30}{1}*\binom{30}{2}*\binom{30}{3}*\binom{30}{1}*\binom{30}{1}*\binom{30}{1} \approx$ 47.7 billion possible lineups. It would take far too long to examine the total value of every single one of these possible lineups and to find the lineup with the highest value under the salary cap.

Thus in order to optimize for value, we need an algorithm that will converge to the optimal lineup

most of the time and that will do so efficiently.

The task of taking a set of objects with values and weights (costs) and finding the combination of objects with the highest total value while keeping the total cost under a certain threshold is commonly referred to as the "knapsack problem."[Mis12] In the classic version of the knapsack problem, you have a set of objects, each with a weight and a value, and you need to fill a bag with the collection of objects that have the greatest total value while have the total weight be under a certain threshold. [Mis12] This is slightly different from the problem of creating an optimal fantasy lineup since the fantasy lineup contains a set number of players from a set number of categories.

There are many different approaches to solving the classic knapsack problem. One method uses a process called "simulated annealing." An outline of the simulated annealing algorithm is below:

Knapsack Algorithm With Simulated Annealing:[Mis12]

Step One: Pick an object randomly (all objects have equal probability of being selected) and add it to the bag

Step Two: If the bag is still under the weight limit, proceed to step three. If the bag is now over the weight limit, randomly select an object in the bag and remove it (all objects have equal probability of being selected). Keep randomly removing objects until the bag is back under the weight limit and then proceed to step three.

Step three: Record the value of the current bag and compare it to the value of the previous bag. If the previous bag has a higher value than the current bag, then make the current bag the new starting bag with probability p and go back to step one. Otherwise reject the current bag and use the previous bag as the new starting bag and go back to step one.

In the simulated annealing version of this algorithm, p is a function of time and decreases over time. This essentially means that in the first several iterations, the algorithm is more willing to explore bags with lower values and thus is less likely to get trapped in a local optimum. Over time, the algorithm becomes less and less willing to explore bags with lower values and thus hopefully eventually converges to the global optimum.

For daily fantasy football lineups, each player's "weight" is his salary, and each players "value" could be his expected points, ceiling, floor, or any other metric that pertains to his performance.

Picking an optimized starting daily fantasy football lineup represents a slightly different challenge than the classical knapsack problem, however, because of the constraints on the number of players from each position that have to be in the lineup.

The R software, lpsolve.api is capable of solving this modified knapsack problem using an algorithm called the branch-and-bound algorithm. This algorithm essentially encodes potential solutions in a tree-structure and then explores branches of the tree, pruning all branches that do not satisfy the constraints of the problem or that have no chance of yielding a solution that is better than the current optimum.[bra] This algorithm thus operates very similarly to the knapsack algorithm, but explores potential solutions in a much more structured way.

4.3 Choosing "Value"

While much work has gone into determining solutions for the knapsack problem in general and into creating programs that optimize fantasy lineups based on expected points, to the best of my knowledge no studies so far have investigated the choice of the "value" that is plugged into these fantasy football optimization programs, despite the fact that many fantasy football blogs and information centers talk about the various player attributes other than expected points, as discussed above in the player terminology section.[Spr15]

There are countless other variables such as ceiling, floor, or consistency, that could be used as the "value" that is optimized in the modified-knapsack problem. This thesis attempts to start to lay the groundwork for a comprehensive analysis and comparison of optimization strategies focused on player attributes other than simply expected points.

This type of analysis is particularly relevant given the variety of contest structures available in daily fantasy football. For example, as mentioned previously, in a 50/50 contest, when all you have to do is beat half the field, it probably makes sense to optimize for attributes that correspond with consistent performance just above average. Meanwhile in a tournament contest, in which it pays to have the highest-scoring lineup possible, it probably makes sense to optimize for player attributes that lead to lineups with higher variances and higher ceilings.

4.4 Correlation between Players

Another commonly recommended daily fantasy football strategy in tournaments is to have some correlation between players.[4fo16] If several offensive players from the same team are in your lineup, then if one of them does well, it probably means that the offense as a whole is doing well, and thus the other players from your lineup who are on that team are doing well. This is particularly the case if one of the players is a quarterback and another is a wide receiver or a tight end on the same team since every time the wide receiver or tight end catches a touchdown, both that player and your quarterback will be earning points. Having two players on the same team is not always a good thing, however, especially if they play the same position. For example if you have two running backs on the same team and there is a situation where the team is attempting to score a rushing touchdown, only one of your players can get points as both players essentially "steal" touches from each other. For this reason, daily fantasy users need to be careful to look for positive correlation in their lineups as opposed to negative correlation if they hope to have a high-scoring team.

While player correlations is a fascinating topic in daily fantasy football and should be an important factor when considering lineup construction, for the scope of this thesis I chose to focus purely on lineups that did not have high player correlations and thus restricted the number of players in the lineup on any given team to one.

5 Data and Methodology

Figure 3: Empirical Fantasy Point Distribution Statistics By Position

Position	Mean	Standard Deviation
Quarterback	13.16	8.80
Running Back	7.37	7.73
Wide Receiver	6.97	7.02

Most of the attributes that I consider optimizing for in lineup construction, such as a player's floor or ceiling, are variables that correspond to a player's fantasy points probability distribution for the game in question. For example, a running back with a high ceiling for a given game might have a greater than 10% chance of putting up 25 or more fantasy points in that game. Meanwhile a running back with a high floor for a given game might have a less than 10% chance of scoring fewer than 8 fantasy points. The choice of what exactly constitutes a high floor or ceiling is a decision

for the optimizer to make, but whatever they decide will be a reflection about the shape of the estimated points distribution for the player in question.

For this reason, my research for this thesis had three phases: building a model that estimates each player's conditional fantasy point distribution for any given game; creating quantitative representations of player attributes based on these conditional distributions; and assessing the relative performance of various lineup optimization strategies using each of these attributes as the "value" variable.

5.1 Building Estimated Points Distributions

5.2 Obtaining Player Data

For the scope of this thesis, I chose to use FanDuel data[gur] since FanDuel is one of the older daily fantasy football platforms and thus had the largest amount of available historical player salary data. I was able to get FanDuel salary data from the 2011 NFL season onwards, and so I focused my data collection on retrieving statistics for players who played in any of the 2011 through 2017 NFL regular seasons.

Most fantasy football layouts including FanDuel's involve building lineups out of a certain number of quarterbacks, running backs, wide receivers, tight ends, kickers, and defenses. For the scope of this thesis, however, I chose to simply focus on quarterbacks, running backs, and wide receivers. This was because modeling player performance for kickers and defenses presented slightly different challenges than modeling performance for offensive skill players like quarterbacks, running backs, and wide receivers. Furthermore, I did not include tight ends in my analysis since the tight end data I had access to was very inconsistently formatted, to the point of making it impossible to scrape player data other than on a one-by-one basis.

That said, while my model is not a full model for every position typically seen in daily fantasy football lineups, my approach and methods are definitely applicable to the whole spectrum of positions and are definitely worth exploring at some point in the future.

I obtained player data by scraping regular season game logs for the top one-hundred running backs, quarterbacks, and tight ends in terms of games played for each regular season from 2011 through 2017. This scraping resulted in obtaining data on a grand total of 145 unique quarterbacks,

220 unique running backs, and 273 unique wide receivers. The data came from pro-football-reference.com[pro], an online database containing football statistics for every team and player for every NFL season since 1922.

Player game logs include some general information about each game as well as the players individual statistics in all categories in the game. For example below is a subset of Pittsburgh Steelers running back Le'Veon Bell's game log:

Figure 4: Le'Veon Bell 2017 Regular Season Game Log

									Rushing				Receiving						
Rk	Year ▼	Date	G#	Age	Tm		Opp	Result	Att	Yds	Y/A	TD	Tgt	Rec	Yds	Y/R	TD	Ctch%	Y/Tgt
1	2017	2017-09-10	1	25-204	PIT	@	CLE	W 21-18	10	32	3.20	0	6	3	15	5.00	0	50.0%	2.50
2	2017	2017-09-17	2	25-211	PIT		MIN	W 26-9	27	87	3.22	0	4	4	4	1.00	0	100.0%	1.00
3	2017	2017-09-24	3	25-218	PIT	@	CHI	L 17-23	15	61	4.07	1	7	6	37	6.17	0	85.7%	5.29
4	2017	2017-10-01	4	25-225	PIT	@	BAL	W 26-9	35	144	4.11	2	6	4	42	10.50	0	66.7%	7.00
5	2017	2017-10-08	5	25-232	PIT		JAX	L 9-30	15	47	3.13	0	10	10	46	4.60	0	100.0%	4.60
6	2017	2017-10-15	6	25-239	PIT	@	KAN	W 19-13	32	179	5.59	1	6	3	12	4.00	0	50.0%	2.00
7	2017	2017-10-22	7	25-246	PIT		CIN	W 29-14	35	134	3.83	0	3	3	58	19.33	0	100.0%	19.33
8	2017	2017-10-29	8	25-253	PIT	@	DET	W 20-15	25	76	3.04	1	3	2	5	2.50	0	66.7%	1.67
9	2017	2017-11-12	9	25-267	PIT	@	IND	W 20-17	26	80	3.08	0	6	5	32	6.40	0	83.3%	5.33
10	2017	2017-11-16	10	25-271	PIT		TEN	W 40-17	12	46	3.83	0	11	9	57	6.33	0	81.8%	5.18
11	2017	2017-11-26	11	25-281	PIT		GNB	W 31-28	20	95	4.75	0	14	12	88	7.33	0	85.7%	6.29

From this game log, we can see that in week one of the 2017 season, Le'Veon Bell played for the Pittsburgh Steelers (PIT) against the Cleveland Browns (CLE) and put up 32 rush yards and 0 rush touchdowns on 10 attempts while hauling in 3 receptions for 15 receiving yards on 6 targets. The format of these game logs is different for wide receivers, running backs, quarterback, but they all contain similar game information and player statistics.

Below is a description of the variables included in quarterback, running back, and wide receiver game logs:

Variable Type	Variables
Game	Year, Date, Game Number, Player's Age, Team, Home/Away, Opponent, Result
Passing	Completions, Attempts, Yards, Touchdowns, Interceptions
Rushing	Attempts, Yards, Touchdowns
Receiving	Targets, Receptions, Yards, Touchdowns

5.3 Adding Fantasy Points

Since fantasy points are not included in pro-football-reference gamelogs, I had to compute the fantasy points earned by each player in each game using the following formula:

$$\begin{aligned} FantasyPoints = {} & PassingYards * 0.25 + PassingTDs * 4 + Interceptions * -1 \\ & + RushingYards * 0.1 + RushingTDs * 6 \\ & + ReceivingYards * 0.1 + ReceivingTDs * 6 \\ & + Fumbles * -2 + TwoPointConversions * 2 \end{aligned} \tag{1}$$

This is the same formula that FanDuel uses for Half-PPR contests, and is the same formula that they have used while generating the salary data for players in each position for the last six years.[fan]

5.4 Adding Player Averages Over a Number of Time Spans

In order to help predict player performance in a given game, I chose to compute summary statistics for each player's career fantasy performance and his recent performance leading leading up to the game in question.

Figure 5: Player Average Fantasy Point Variables

Variable	Description
FPoints_A	The player's average fantasy points in the his entire career leading up to the current game.
FPoints_4	The player's average fantasy points in the four games immediately prior to the current game.
FPoints_10	The player's average fantasy points in the ten games immediately prior to the current game.

These three variables essentially indicate a measure of player quality over three different time frames: career, 10-game span, and 4-game span. I decided to include the shorter time frame averages because the NFL is a rapidly changing league and even if a player's individual characteristics do not change much over a couple seasons, the teams they play for have significant roster turnover year to year (usually 15%-40%), and and the roles that they serve on these teams can change significantly even within the same season.

5.5 Adding Game Data

While the player game logs contain lots of relevant variables, such as players' rushing attempts, yards, touchdowns, and many others, they do not include any other information on the game or on the defense that the player is matching up against. For this reason, I supplemented the player game-logs from pro-football-reference.com with additional game-level data for each game in the game-log.

Below is some game specific data for the Steelers' week one game in 2017 against the Cleveland Browns:

	1	2	3	4	Final
Pittsburgh Steelers	7	7	7	0	21
Cleveland Browns	7	0	3	8	18

As we can see, Le'Veon Bell's week one game started at 1:00pm at the FirstEnergy Stadium and had 67,431 people in attendance. Other game data of interest that is available before the start of the game can be found in the game info boxes such as the one below. These boxes are on the game page for each game on pro-football-reference.com.

Game Info	
Won Toss	Steelers (deferred)
Roof	outdoors
Surface	grass
Weather	63 degrees, wind 11 mph
Vegas Line	Pittsburgh Steelers -10.0
Over/Under	47.0 **(under)**

From this game info box, we can see that this Steelers-Browns game was played outdoors on grass in 63 degrees with winds of 11 MPH. The Las Vegas betting line going into the game was Steelers -10.0, and the over/under line was 47.0. For those unfamiliar with sports betting, a Vegas line of Steelers -10.0 means that bettors who want to bet on the line can either bet on the Steelers scoring 10+ more points than the Browns or bet on the Steelers scoring fewer than 10 more points than the Browns. Meanwhile, the over/under line is a statement about the total number of points scored by both teams during the course of the game. If both teams combine to score more than 47.0 points, then this game will have gone "over," whereas if they combine to score fewer than 47.0 points, the game will have gone "under." As with the Vegas line, bettors can place bets on the game going over or under the over/under line.

5.6 Adding Defense Data

While general game data and player-specific data are useful, the quality of the defense that player is going against in the current week is also very important in predicting player performance. Some defenses are capable of shutting down offenses and limited opposing players to sub-par fantasy performances, while other defenses are incapable of getting a stop and can give up huge statistical performances to opposing players. For this reason, I also scraped data on every defense for every game in every player's game-logs.

To do this I first created a database of defense game-logs for each team for each season from 2011 to 2017.

Below is an example defense game-log for the 2015 Seattle Seahawks:

										Scor	Scor	Defe	Defe	Defe	Defe	Defe
Week	Day	Date				OT	Rec		Opp	Tm	Opp	1stD	TotYd	PassY	RushY	TO
1	Sun	September 13	1:00PM ET	boxscore	L	OT	0-1	@	St. Louis Rams	31	34	19	352	276	76	3
2	Sun	September 20	8:25PM ET	boxscore	L		0-2	@	Green Bay Packers	17	27	21	361	234	127	1
3	Sun	September 27	4:26PM ET	boxscore	W		1-2		Chicago Bears	26	0	7	146	48	98	
4	Mon	October 5	8:31PM ET	boxscore	W		2-2		Detroit Lions	13	10	12	256	203	53	1
5	Sun	October 11	1:03PM ET	boxscore	L	OT	2-3	@	Cincinnati Bengals	24	27	27	419	310	109	2
6	Sun	October 18	4:06PM ET	boxscore	L		2-4		Carolina Panthers	23	27	25	383	248	135	2
7	Thu	October 22	8:26PM ET	boxscore	W		3-4	@	San Francisco 49ers	20	3	8	142	81	61	
8	Sun	November 1	4:25PM ET	boxscore	W		4-4	@	Dallas Cowboys	13	12	14	220	91	129	
9									**Bye Week**							
10	Sun	November 15	8:35PM ET	boxscore	L		4-5		Arizona Cardinals	32	39	30	451	334	117	3
11	Sun	November 22	4:27PM ET	boxscore	W		5-5		San Francisco 49ers	29	13	14	306	247	59	
12	Sun	November 29	4:26PM ET	boxscore	W		6-5		Pittsburgh Steelers	39	30	26	538	480	58	4
13	Sun	December 6	1:02PM ET	boxscore	W		7-5	@	Minnesota Vikings	38	7	9	125	94	31	1
14	Sun	December 13	1:02PM ET	boxscore	W		8-5	@	Baltimore Ravens	35	6	16	302	274	28	2
15	Sun	December 20	4:06PM ET	boxscore	W		9-5		Cleveland Browns	30	13	15	230	136	94	1
16	Sun	December 27	4:26PM ET	boxscore	L		9-6		St. Louis Rams	17	23	14	205	103	102	
17	Sun	January 3	4:25PM ET	boxscore	W		10-6	@	Arizona Cardinals	36	6	16	232	205	27	3

Using these defense game-logs, I was then able to compute a defenses average statistics on the year going into each game of the season. So for every offensive player, I appended the year to date averages for the defense that they matched up against each week in several different variables. I only included year-to-date data because defenses change significantly year to year due to trades, injuries, and other circumstances that lead to personnel changes.

Figure 6: Key Defense Variables

Variables	Description
DPtsA	Average Points Against
DFDA	Average First Downs Against
DPassYdsA	Average Passing Yards Against
DRushYdsA	Average Rushing Yards Against
DTdsA	Average Touchdowns Against

5.7 OLS Regressions

After collecting all the player game-logs and supplementing them with additional game data, defense data, and player fantasy points averages, I then fit three different regression models—one for quarterbacks, one for running backs, and one for wide receivers—for predicting the amount of fantasy points a player will earn in a given week.

While these models are interesting in and of themselves, I fit them in order to see the relative effects of each predictor variable on predicting a player's performance. As explained in the section below, this information is critical for determining how to appropriately weight observations when creating a player's estimated conditional fantasy points distribution for a given game.

I fit different models for each position because each predictor variable affects each position in a different way. For example, if a defense gives up a lot of rushing yards, that will have more of an impact on the running backs who go up against them than on the wide receivers who face them. Similarly, high winds may negatively impact quarterbacks' fantasy points by making it harder to throw the ball down field, but might have no effect on running backs' fantasy points.

I chose to perform OLS regression because for there seems to be a linear relationship between fantasy points production and most of the binary variables for each position. Furthermore, model diagnostic tests show that there does not appear to be significant evidence of heteroskedasticity or of non-normality of the error terms in any of the three models.

Here is the final regression model for quarterbacks after conducting step-wise backwards regression:

Dep. Variable:	FPoints	**R-squared:**	0.454
Model:	OLS	**Adj. R-squared:**	0.454
Method:	Least Squares	**F-statistic:**	712.7
Date:	Mon, 02 Apr 2018	**Prob (F-statistic):**	0.00
Time:	12:27:34	**Log-Likelihood:**	-25298.
No. Observations:	7720	**AIC:**	5.062e+04
Df Residuals:	7710	**BIC:**	5.069e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	-10.3170	0.706	-14.620	0.000	-11.700 -8.934
Home	0.5787	0.146	3.960	0.000	0.292 0.865
G_Num	-0.0372	0.016	-2.332	0.020	-0.068 -0.006
Wind	-0.0774	0.015	-5.114	0.000	-0.107 -0.048
Age	-0.0739	0.019	-3.876	0.000	-0.111 -0.037
FPoints_A	0.4243	0.025	16.739	0.000	0.375 0.474
FPoints_4	0.1482	0.017	8.685	0.000	0.115 0.182
DPtsA	0.2113	0.016	13.012	0.000	0.179 0.243
DPassYdsA	0.0305	0.002	14.583	0.000	0.026 0.035
Start	9.1577	0.250	36.684	0.000	8.668 9.647

Omnibus:	202.758	**Durbin-Watson:**	1.794
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	229.852
Skew:	0.367	**Prob(JB):**	1.23e-50
Kurtosis:	3.420	**Cond. No.**	2.28e+03

Variable Interpretations:

Home: Being at home increases a QB's predicted points by 0.58.

G_Num: QBs are predicted to score 0.037 fewer points in each successive game week over the course of the season.

Wind: QBs are predicted to score 0.8 fewer points per each additional 10mph of wind.

Age: QBs are predicted to score 0.07 fewer points each additional year they age.

FPoints_A: Qbs are predicted to score 0.42 more points for each additional point in their career fantasy points average.

FPoints_4: Qbs are predicted to score 0.15 more points for each addition point in their fantasy points average over the last four games.

DPtsA: Qbs are predicted to score 0.21 more points for each additional point the defense they face has given up on average over the course of the season.

DPassYdsA: Qbs are predicted to score 3.1 more points for each additional 100 passing yards the defense they face has given up on average over the course of the season.

Start: Qbs are predicted to score 9.16 more points if they start.

Here is the final regression model for estimating running backs fantasy points after conducting step-wise backwards regression:

Dep. Variable:	FPoints	**R-squared:**	0.418
Model:	OLS	**Adj. R-squared:**	0.418
Method:	Least Squares	**F-statistic:**	733.4
Date:	Sun, 01 Apr 2018	**Prob (F-statistic):**	0.00
Time:	13:55:25	**Log-Likelihood:**	-36025.
No. Observations:	11238	**AIC:**	7.207e+04
Df Residuals:	11226	**BIC:**	7.216e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	-0.3015	0.746	-0.404	0.686	-1.763 1.160
Home	0.6579	0.113	5.835	0.000	0.437 0.879
Temp	-0.0101	0.004	-2.576	0.010	-0.018 -0.002
Age	-0.2530	0.022	-11.479	0.000	-0.296 -0.210
FPoints_A	0.4628	0.023	19.951	0.000	0.417 0.508
FPoints_4	0.2696	0.020	13.598	0.000	0.231 0.309
FPoints_10	0.0516	0.026	1.992	0.046	0.001 0.102
DPtsA	0.0817	0.015	5.560	0.000	0.053 0.111
DFDA	0.0765	0.028	2.709	0.007	0.021 0.132
DRushYdsA	0.0354	0.003	13.857	0.000	0.030 0.040
Surface	0.4073	0.116	3.503	0.000	0.179 0.635
Start	3.0940	0.135	22.835	0.000	2.828 3.360

Omnibus:	2063.306	**Durbin-Watson:**	1.633
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4402.846
Skew:	1.078	**Prob(JB):**	0.00
Kurtosis:	5.180	**Cond. No.**	1.77e+03

Variable Interpretations:

Home: RBs are expected to score 0.658 more points if they are playing at home.

Temp: RBs are expected to score 0.1 fewer points for each additional 10 degrees of temperature.

Age: RBs are expected to score 0.25 fewer points for each additional year that they age.

FPoints_A: RBs are expected to score 0.46 more points for every point higher their career fantasy points average is.

FPoints_4: RBs are expected to score 0.27 more points for every point higher their fantasy points average over the last four games is.

FPoints_10: RBs are expected to score 0.05 more points for every point higher their fantasy points average over the last ten games is.

DPtsA: RBs are expected to score 0.8 more points for each additional 10 points the defense they are facing has given up on average that season.

DFDA: RBs are expected to score 0.7 more points for each additional 10 first downs the defense they are facing has given up on average that season.

DRushYdsA: RBs are expected to score 3.5 more points for each additional 100 yards of rushing the defense they are facing has given up on average that season.

Surface: RBs are expected to score 0.4 more points when playing on turf.

Start: RBs are expected to score 3.1 more points when they start.

Final wide receivers step-wise regression model:

Dep. Variable:	FPoints	**R-squared:**	0.306
Model:	OLS	**Adj. R-squared:**	0.305
Method:	Least Squares	**F-statistic:**	805.4
Date:	Mon, 02 Apr 2018	**Prob (F-statistic):**	0.00
Time:	12:37:26	**Log-Likelihood:**	-58535.
No. Observations:	18305	**AIC:**	1.171e+05
Df Residuals:	18294	**BIC:**	1.172e+05
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	0.6310	0.521	1.211	0.226	-0.391 1.653
G_Num	-0.0428	0.009	-4.505	0.000	-0.061 -0.024
Age	-0.2186	0.015	-14.154	0.000	-0.249 -0.188
FPoints_A	0.5237	0.022	23.711	0.000	0.480 0.567
FPoints_4	0.1806	0.018	10.119	0.000	0.146 0.216
FPoints_10	0.1874	0.024	7.887	0.000	0.141 0.234
DPtsA	0.0538	0.011	5.010	0.000	0.033 0.075
DPassYdsA	0.0170	0.001	13.448	0.000	0.015 0.019
DTdsA	0.2315	0.075	3.070	0.002	0.084 0.379
Dome	-0.4366	0.099	-4.424	0.000	-0.630 -0.243
Start	1.8421	0.105	17.622	0.000	1.637 2.047

Omnibus:	3362.074	**Durbin-Watson:**	1.816
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	6829.225
Skew:	1.102	**Prob(JB):**	0.00
Kurtosis:	5.024	**Cond. No.**	2.86e+03

Variable Interpretations:

G_Num: WR's are predicted to score 0.04 fewer points for every additional game over the course of the season.

Age: WR's are predicted to score 0.21 fewer points for every additional year that they are older.

FPoints_A: WR's are predicted to score 0.52 more points for every additional point in their career fantasy points average.

FPoints_4: WR's are predicted to score 0.18 more points for every additional point in their fantasy points average over the last four games.

FPoints_10: WR's are predicted to score 0.19 more points for every additional point in their fantasy points average over the last ten games.

DPtsA: WR's are predicted to score 0.5 more points for every additional 10 points the defense

they are facing has given up on average over the course of the season.

DPassYdsA: WR's are predicted to score 1.7 more points for every additional 100 yards of passing the defense they are facing has given up on average over the course of the season.

DTdsA: WR's are predicted to score 0.23 more points for every additional touchdown the defense they are facing has given up on average over the course of the season.

Dome: WR's are predicted to score 0.44 fewer points when playing under a dome.

Start: WR's are predicted to score 1.84 more points when starting.

5.8 Generating Estimated Conditional Fantasy Points Distributions

5.8.1 Determining Observation Weights

When estimating a player's conditional fantasy points distribution for a given game, ideally we would get the empirical conditional distribution simply by subsetting the player's game data and looking at their fantasy points distribution in games where all the predictor variables have exactly same values as they do for the game in question.

However, since each game is relatively unique, there are often no prior games in a player's career where all the predictor variables perfectly match up with those in the current game. In other words, a player never players the same game twice since players and teams change over time. That said, there are many games over the course of a player's career which, while not the same, can share many similarities. For example, a player might have two games in similar weather conditions against a similar caliber defense and with similar caliber offensive players around them, etc. Thus the similarity of predictor variables for one game to those of another game can help give us insight into how the player will perform in the game yet to be played.

In order to estimate a player's conditional fantasy points distribution for a game, I decided to use the fantasy points in all the games in the last three years of that player's career leading up to the current game as observations to be put into the distribution, but weighted these observations based on how similar the predictor variables were to those of the game in question and based on how much those predictor variables mattered in predicting a player's performance.

In order to determine how similar the predictor variables are for two distinct games, I first thought to take the sum of distances between variables. So if game a has predictor variables with values $X_{1_a}, X_{2_a}, \ldots, X_{n_a}$, and game b has predictor variables with values $X_{1_b}, X_{2_b}, \ldots, X_{n_a}$, the sum of the distances between the two is $|X_{1_a} - X_{1_b}| + |X_{2_a} - X_{2_b}| + \ldots + |X_{n_a} - X_{n_b}|$. With this setup, two games that are similar with have a sum of distances that is close to 0, whereas two games that are very different will have a sum of distances that is equal to some positive number that is distant from 0.

In order for this type of calculation to be effective, all of the predictor variables have to be on the same scale otherwise a variable like attendance, which is on the order of tens of thousands, will drown out the effects of a variable on the scale of 0-10, like defensive average touchdowns against.

Thus before computing observation weights, I first z-scored all of the non-binary variables that were significant in the OLS regressions for each position. The z-score of a variable has the following formula: $zscore(x_i) = \frac{x_i - \bar{x}}{s}$ where $\bar{x}$ is the observed mean for a predictor variable, and s is the observed standard deviation. The nice thing about z-scored variables is that the z-score of any non-binary variable is approximately normally distributed with mean 0 and variance 1. Thus z-scoring all of my non-binary predictor variables effectively put them all on the same scale.

Under this process, two games that are very similar will have a distance of close to zero between each pair of predictor variables. Similarly, since 95% of the data should fall within 1.96 standard deviations of the mean for normally distributed variables, the distance between two z-scored variables will be less than $2 * 1.96 = 3.92$ most of the time.

In order to put all of my binary predictor variables on the same scale as the z-scored variables, transformed them by multiplying their 0 and 1 values by 3.92. So all binary variables now had values of either 0 or 3.92. Thus if the same binary predictor variable had two different values for two games, the distance between the variables would be 3.92, while if the same binary predictor variables had the same value for two different games, the distance between the variables would be 0.

Now that all of the predictor variables were on approximately the same scale, I decided that the weight of an observation should be the sum of the weights contributed by each of the predictor variables, where the weight that each predictor variable contributes is calculated as follows:

$$W(X_{i_a}) = S(X_{i_a}, X_{i_b}) * c_{X_i}$$

Note that here I am trying to compute the amount that the variable X_i should contribute to the total observation weight of game a when creating the estimated conditional fantasy points distribution for game b.

Formula Breakdown:

1. $W(X_{i_a})$ is the amount that variable X_i should contribute to the total observation weight of game a.
2. $S(X_{i_a}, X_{i_b})$ is a similarity score for the variable X_i where $S(X_{i_a}, X_{i_b}) = \frac{3.92+1.0}{|X_{i_a} - X_{i_b}|+1.0}$.
3. c_{X_i} is the normalized regression coefficient for the variable X_i in the regression model for players with the same position as the player in question. Note that this model is slightly different than the ones calculated in the previous section, since it is based off of the transformed values of the predictor variables (through z-scoring and changing binary variables).

Breakdown of the similarity score:

$S(X_{i_a}, X_{i_b}) = \frac{3.92+1.0}{|X_{i_a} - X_{i_b}|+1.0}$ is a reflection of how similar the variable X_i is for games a and b. If the two variables are exactly the same, then the similarity score will be 4.92. If the two variables are very dissimilar, then $|X_{i_a} - X_{i_b}|$ will be approximately 3.92, so the similarity score will around 1.

Thus if game a is very dissimilar to game b, $\sum_{i=1}^{n} W(X_{i_a}) * c_{X_i} \approx \sum_{i=1}^{n} 1 * c_{X_i} = 1$ since the regression coefficients are normalized.

Choosing to make even games which are very dissimilar to the game for which we are trying to estimate a player's conditional fantasy points distribution have an observation weight of 1 makes sense since every game gives us some information on a player. Off course z-scored variables can range beyond ± 1.96, so the difference in z-scored variables could lead to a total observation of less than 1 in extreme cases. For this reason, I set the observation weight as $\max(\sum W(x_i) * c_{X_i}, 1)$ to ensure that every observation was included in the estimated conditional distribution with at least a weight of 1.

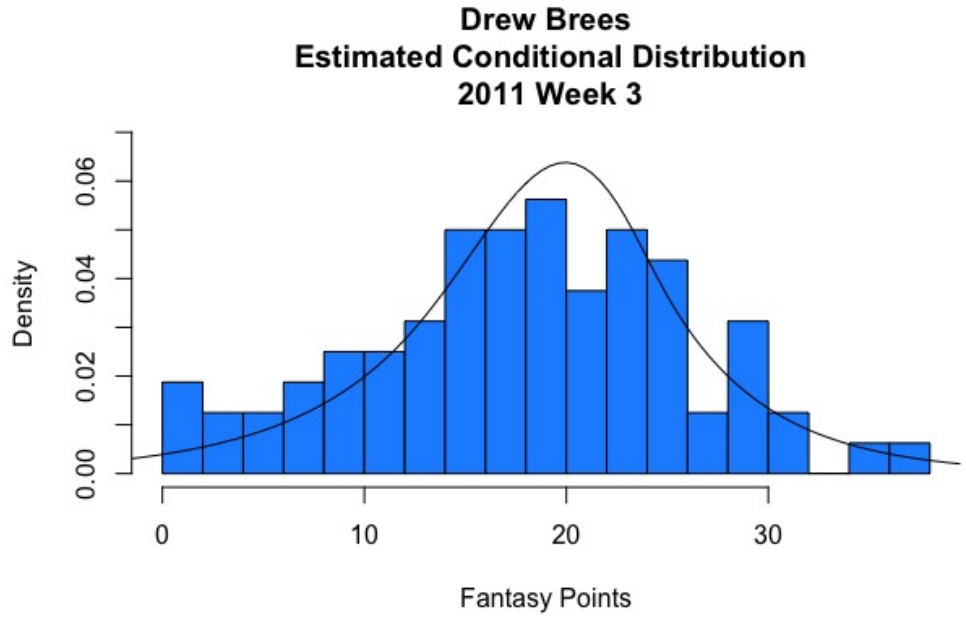
This formula as a whole makes sense since the more games there are in the data that are very similar to the game in question, the more the data from similar games will be represented in the

estimated conditional distribution. Thus as the quantity of data increases, the player's estimated conditional distribution will approximate the player's true conditional distribution.

After weighting each observation, I recorded the observed fantasy points for each observation, weighted according to the observation weights, and incorporated them as data points in the estimated conditional fantasy points distribution.

After generating this set of data points representing an estimated conditional fantasy points distribution for every game for every player, I used R's logspline package to fit logsplines on each dataset as an estimate of the continuous density function for the estimated conditional fantasy distribution for that game.

For example, we can see in the figure below the logspline fit on the histogram of Drew Brees' estimated conditional distribution for week three of the 2011 season in which he scored 24.7 fantasy points:



The logspline fit is based on a maximum likelihood approach of estimating the density function of a given empirical distribution and typically fits the data relatively well.

5.9 Quantifying Fantasy Point Distribution Attributes

As mentioned in a previous section, there are many different player attributes that are of interest when constructing optimized fantasy lineups. The table below outlines each of these desirable

attributes and explains how I chose to quantify them in terms of a player's estimated conditional fantasy points distribution.

Figure 7: Distribution Attributes of interest

Attribute	Formulation
Mean	The mean of the distribution
Median	The median of the distribution
Mode	The fantasy points point-value with the highest density function value
Floor	The 10th percentile of the distribution
Ceiling	The 90th percentile of the distribution
Std Dev.	The standard deviation of the distribution

5.10 Optimization Methods

After having calculated the mean, median, mode, floor, ceiling, and variance of each estimated conditional fantasy points distribution for every game for every player in my dataset, I then tested out the effects of optimizing lineups for different sets of attributes.

Since my lineups only consisted of six players (one quarterback, two running backs, and three wide receivers) instead of the nine players that are typically in a FanDuel lineup, I changed the salary cap from the typical FanDuel cap of \$60,000 to $\frac{6}{9} * \$60,000 = \$40,000$.

I considered optimizing lineups for players with the following attributes: mean, median, high-floor, high-ceiling, high-variance, and high mode.

For each of these optimizations strategies, I calculated the top optimized lineup for every week of each regular season from 2011 to 2017. This resulted in a grand total of 119 lineups created by each strategy. I then computed the total actual score of each these lineups for each optimization method and compared the results. I also included an additional lineup for each week that was randomly generated and had salary within \$2,000 of the \$40,000 salary cap.

In addition to evaluating the results of each optimization method using the actual fantasy points scored by each player in each week, I also evaluated the methods by conducting simulations in which I calculated 1000 different scores for each method for each week by randomly sampling from each player's empirical distributions 1000 times and computing the lineup's total fantasy points for the trial.

6 Results

Figure 8: Actual Method Metrics

Method	Mean	St. Dev.
Mean	72.33	18.60
Median	72.33	18.60
Floor	71.72	18.61
Ceiling	67.94	19.12
Mode	65.75	15.53
St. Dev.	64.29	16.40
Random	58.06	14.55

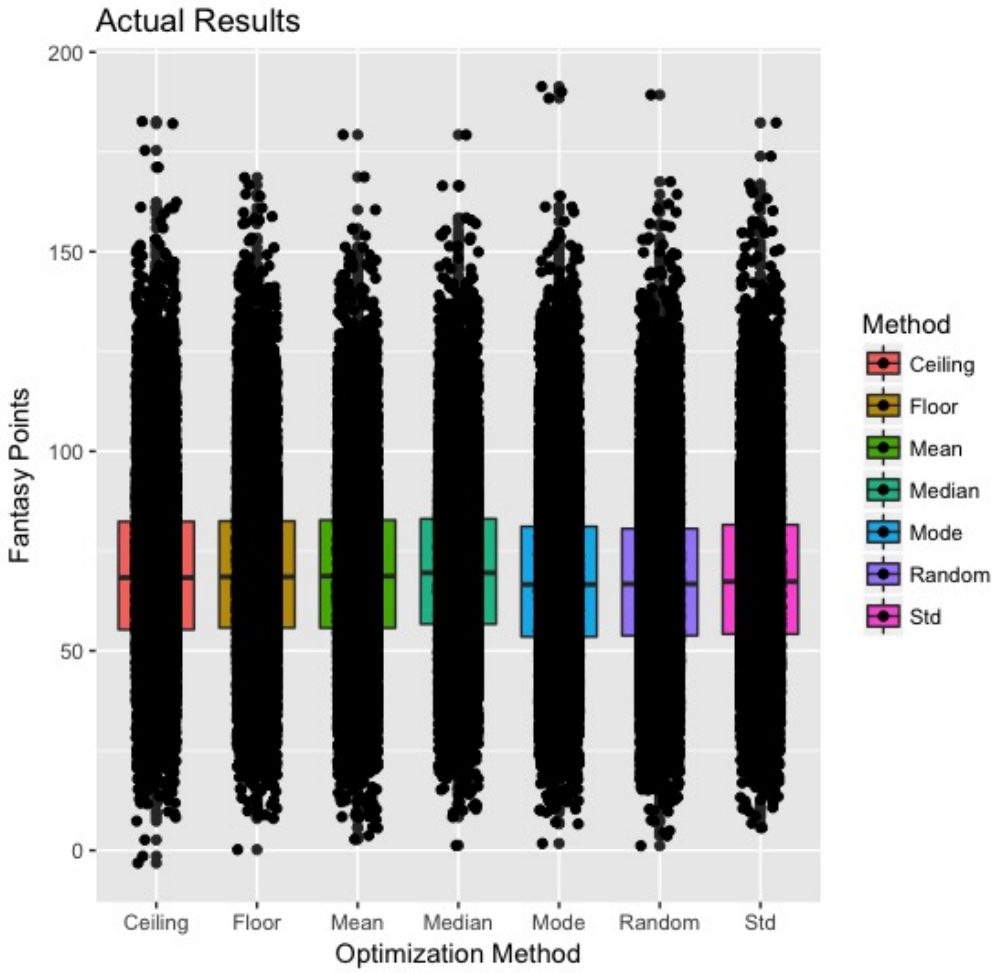
Figure 9: Simulations Method Metrics

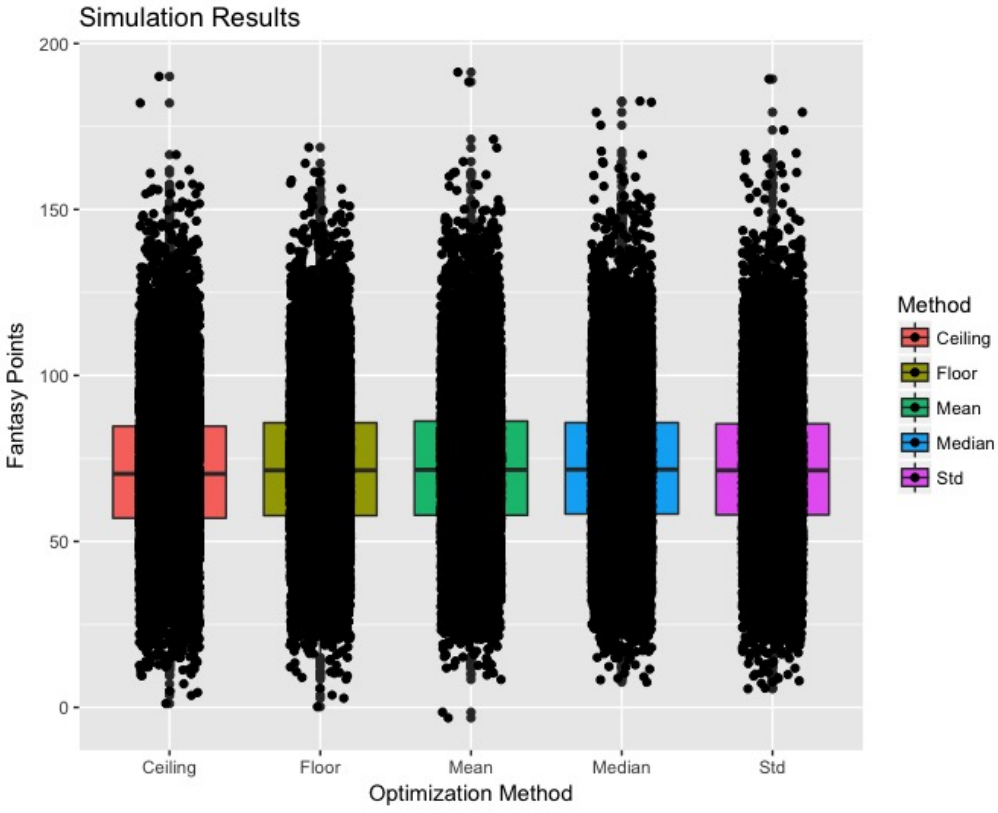
Method	Mean	St. Dev.
Mean	76.31	19.72
Median	76.27	20.15
Floor	72.44	17.95
Ceiling	70.81	23.21
Mode	67.81	16.85
St. Dev.	65.94	23.05
Random	57.06	13.98

Key Takeaways:

1. On average, for both actual scores and the simularions optimizing for the mean, the median, and the floor yields the best results.

2. The algorithm that randomly selected lineups within $2,000 of the salary cap performed the worst on average, a full 14 points below the averages for the mean and median methods


Actual Results
Fantasy Points
200
150
100
50
0
Ceiling
Floor
Mean
Median
Mode
Random
Std
Optimization Method
Method
Ceiling
Floor
Mean
Median
Mode
Random
Std


Simulation Results
Fantasy Points
200
150
100
50
0
Ceiling
Floor
Mean
Median
Std
Optimization Method
Method
Ceiling
Floor
Mean
Median
Std

As we can see from the box plots, while most variables have similar maxes and mins, the for both the simulations and the actual results, the two variables with the greatest number of lineups that scored above 150 are the ceiling optimization and the standard deviation optimization. Meanwhile the optimizations with the fewest amount of lineups with scores below 25 are the floor and median optimizations

Each of these findings supports the notions that higher variance lineups with higher ceilings are likely better for tounament structure, while more conservative optimization strategies are better for 50/50 play.

7 Potential Extensions

While this thesis was a significant undertaking, there remains a lot of work to be done in the analysis of daily fantasy football lineup optimization strategies, particularly with regards to strategies that are not focused purely on projected points.

It would be fascinating to expand this model to include the other positions typically required in fantasy football lineups such as tight ends, kickers, and defenses.

Similarly, since correlation between players is a huge part of daily fantasy strategy, expanding the work done in this paper to include correlations between players is critical in order to more accurately reflect the plethora of lineup construction strategies that daily fantasy users have available to them.

Another possible extension would be to further explore optimizing for other aspects of a players estimated distribution than those studied in this thesis.

Finally, not enough work has been done on examining the effects of not optimizing purely for one type of player in a lineup, but for choosing hybrid lineups that have players with different desirable attributes.

8 Appendix

8.1 Z-Scored Regression Models

Final WR regression model with z-scored data and adjusted binary variables:

Dep. Variable:	FPoints	**R-squared:**	0.305
Model:	OLS	**Adj. R-squared:**	0.305
Method:	Least Squares	**F-statistic:**	893.7
Date:	Sun, 01 Apr 2018	**Prob (F-statistic):**	0.00
Time:	22:25:43	**Log-Likelihood:**	-58539.
No. Observations:	18305	**AIC:**	1.171e+05
Df Residuals:	18295	**BIC:**	1.172e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	6.3313	0.107	58.946	0.000	6.121 6.542
G_Num	-0.1932	0.044	-4.430	0.000	-0.279 -0.108
AgeNew	-0.6735	0.047	-14.291	0.000	-0.766 -0.581
FPoints_A	1.9392	0.082	23.701	0.000	1.779 2.100
FPoints_4	0.9173	0.091	10.089	0.000	0.739 1.096
FPoints_10	0.8576	0.108	7.923	0.000	0.645 1.070
DPtsA	0.2096	0.050	4.158	0.000	0.111 0.308
DPassYdsA	0.7116	0.050	14.187	0.000	0.613 0.810
Dome	-0.1119	0.025	-4.445	0.000	-0.161 -0.063
Start	0.4709	0.027	17.670	0.000	0.419 0.523

Omnibus:	3370.558	**Durbin-Watson:**	1.816
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	6864.761
Skew:	1.103	**Prob(JB):**	0.00
Kurtosis:	5.032	**Cond. No.**	13.5

Final QB regression model with z-scored dataand adjusted binary variables:

Dep. Variable:	FPoints	**R-squared:**	0.440
Model:	OLS	**Adj. R-squared:**	0.440
Method:	Least Squares	**F-statistic:**	674.4
Date:	Sun, 01 Apr 2018	**Prob (F-statistic):**	0.00
Time:	22:28:47	**Log-Likelihood:**	-25394.
No. Observations:	7720	**AIC:**	5.081e+04
Df Residuals:	7710	**BIC:**	5.088e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	6.6631	0.220	30.236	0.000	6.231 7.095
G_Num	-0.1719	0.075	-2.301	0.021	-0.318 -0.025
Wind	-0.3609	0.074	-4.889	0.000	-0.506 -0.216
AgeNew	-0.4067	0.077	-5.250	0.000	-0.559 -0.255
FPoints_A	2.2443	0.125	18.005	0.000	2.000 2.489
FPoints_4	1.1392	0.116	9.845	0.000	0.912 1.366
DPtsA	1.0711	0.085	12.593	0.000	0.904 1.238
DPassYdsA	1.1920	0.086	13.900	0.000	1.024 1.360
Home	0.1501	0.038	3.976	0.000	0.076 0.224
Start	2.0249	0.060	33.537	0.000	1.907 2.143

Omnibus:	219.040	**Durbin-Watson:**	1.748
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	245.426
Skew:	0.392	**Prob(JB):**	5.09e-54
Kurtosis:	3.385	**Cond. No.**	13.2

Final RB regression model with z-scored data and adjusted binary variables:

Dep. Variable:	FPoints	**R-squared:**	0.418
Model:	OLS	**Adj. R-squared:**	0.417
Method:	Least Squares	**F-statistic:**	732.8
Date:	Sun, 01 Apr 2018	**Prob (F-statistic):**	0.00
Time:	22:31:17	**Log-Likelihood:**	-36027.
No. Observations:	11238	**AIC:**	7.208e+04
Df Residuals:	11226	**BIC:**	7.217e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	5.6779	0.111	51.143	0.000	5.460 5.896
AgeNew	-0.6811	0.060	-11.440	0.000	-0.798 -0.564
Temp	-0.1420	0.056	-2.515	0.012	-0.253 -0.031
FPoints_A	2.2243	0.111	19.960	0.000	2.006 2.443
FPoints_4	1.6436	0.121	13.618	0.000	1.407 1.880
FPoints_10	0.2841	0.144	1.973	0.049	0.002 0.566
DPtsA	0.4018	0.078	5.125	0.000	0.248 0.556
DPassYdsA	0.1782	0.072	2.470	0.014	0.037 0.320
DRushYdsA	0.9899	0.069	14.290	0.000	0.854 1.126
Surface	0.1037	0.030	3.494	0.000	0.046 0.162
Home	0.1687	0.029	5.866	0.000	0.112 0.225
Start	0.7851	0.035	22.725	0.000	0.717 0.853

Omnibus:	2063.699	**Durbin-Watson:**	1.633
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4406.235
Skew:	1.078	**Prob(JB):**	0.00
Kurtosis:	5.182	**Cond. No.**	12.0

Aggregrate Plots:

Figure 10: Performance of NFL QBs Who Played Between 2011 and 2017

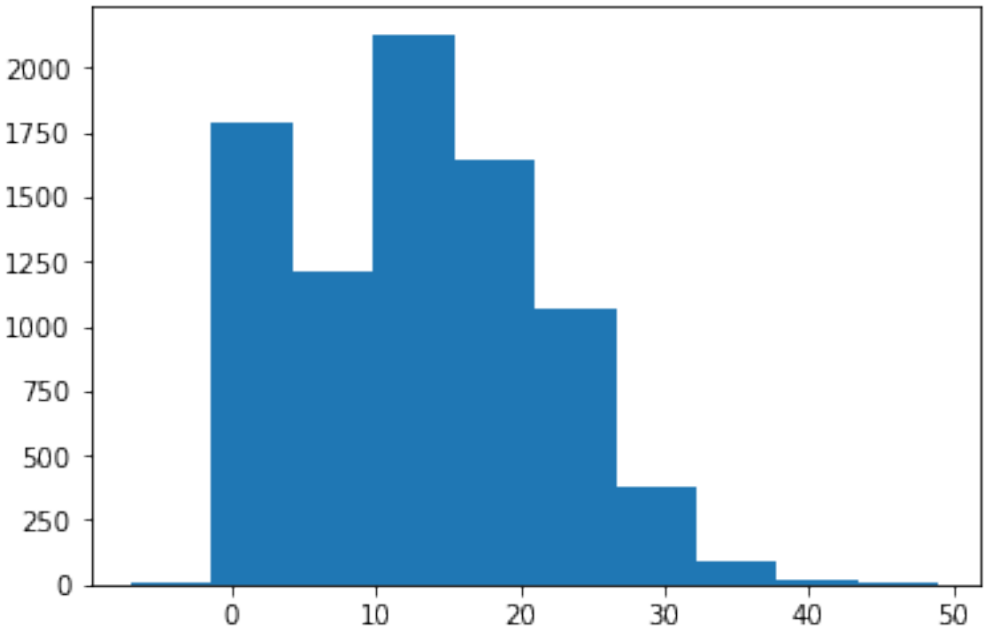


Figure 11: Performance of NFL RBs Who Played Between 2011 and 2017

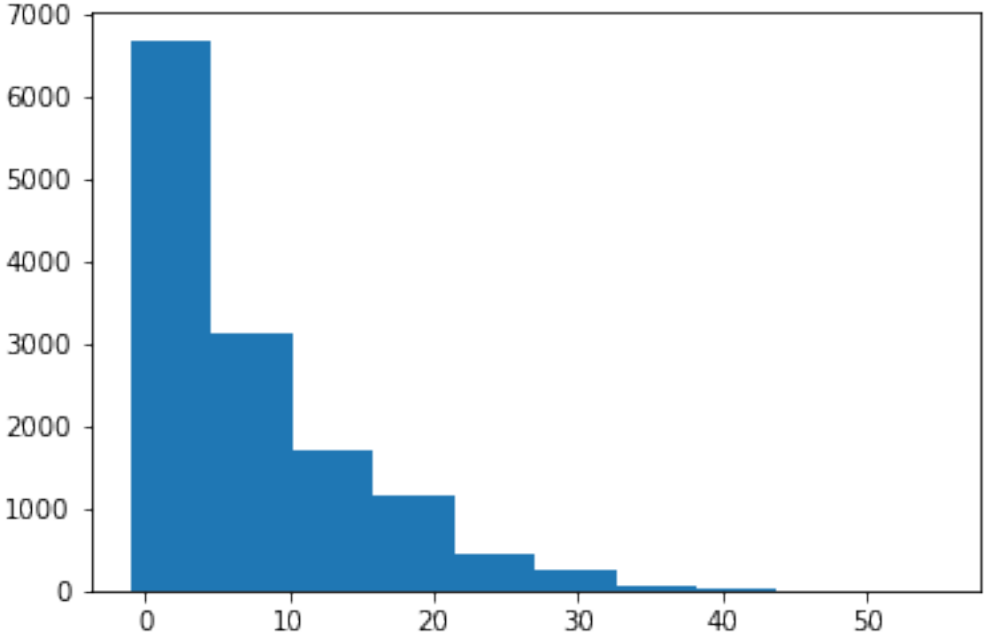
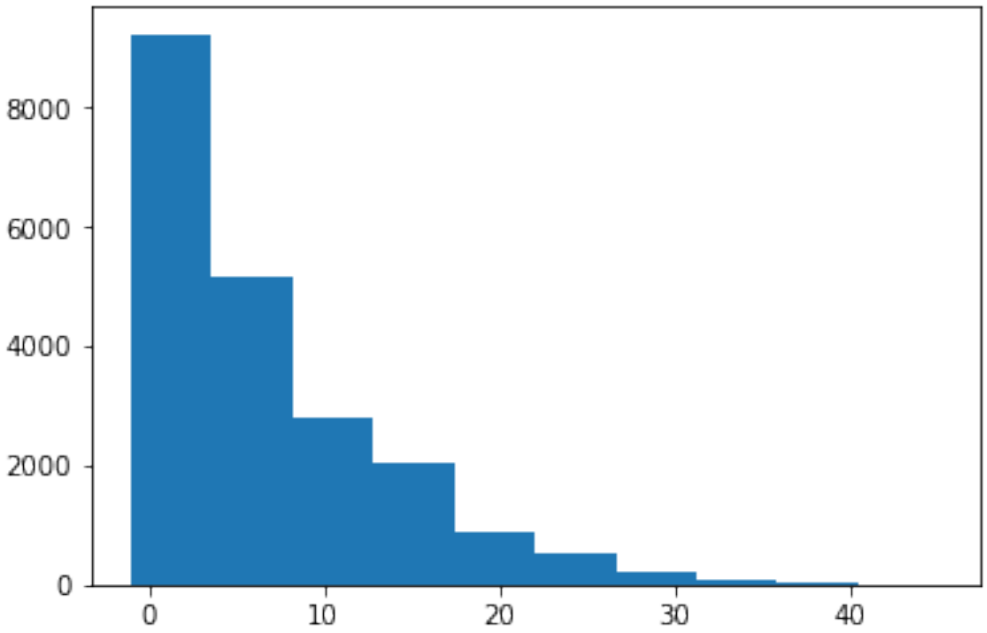


Figure 12: Performance of NFL WRs Who Played Between 2011 and 2017



References

[4fo16] The definitive guide to stacking on draftkings, Aug 2016.

[And16] Kevin Anderton. Fanduel and draftkings are dominating the daily fantasy sports market [infographic]. *Forbes*, Dec 2016.

[bra]

[Bro16] Eddie Brown. The history of fantasy football. *sandiegouniontribune.com*, Sep 2016.

[dfaa] 50/50 multipliers - fanduel - for beginners.

[dfab] Fanduel head to head how does it work? - review.

[dfac] Tournaments fanduel how do they work? - review.

[fan] Rules scoring.

[gur]

[Mis12] Benjamin Misch. Simulated annealing and the knapsack problem, Dec 2012.

[pro] Pro football statistics and history.

[rul] Roster rules refresher: Practices squad, ir, pup.

[Sil17] Steve Silverman. Types of football running backs, Sep 2017.

[Spr15] Jason Spry. Home, Oct 2015.

[Zel] Alex Zelvin. Behind fanduel's salary cap. *FanDuel Fantasy Sports Salary Cap Calculation.*