Fantasy Football Capstone Project

Alex Brandt

December 30, 2018

Introduction

Fantasy football is a competitive game that is played by millions of friends and colleagues every football season. The goal is to score more points that your opponent each week by putting together the best possible team of quarterbacks, running backs, wide receivers, tight ends, kickers, and defense. While there are several different formats, the data in this project utilized PPR (Points per reception) format.

Problem Statement

There are many different factors that can impact your final standing in your league, however, I will be focusing on starting the season off strong by having the best possible draft. I will be using the final 2017 data from Fantasydata.com to try and predict which players have the best value. This data can be used before the season by any potential fantasy league player trying to determine when to draft specific players in order to score the most points in any given week during the season. While adding undervalued players through the waiver wire during the season is an important part to any winning team, this project is only focusing on the initial draft. Will drafting Todd Gurley with the first overall pick help my team make the playoffs? Should I draft Christian Mccaffrey in the first round, or wait until my second pick? These are some of the questions this project will attempt to answer.

I will be narrowing the scope of this project to focus specifically on the running back position. Using the ADP (Average Draft Position) for the 2018 season, I will be able to test the model using the previous year's data.

Data Wrangling

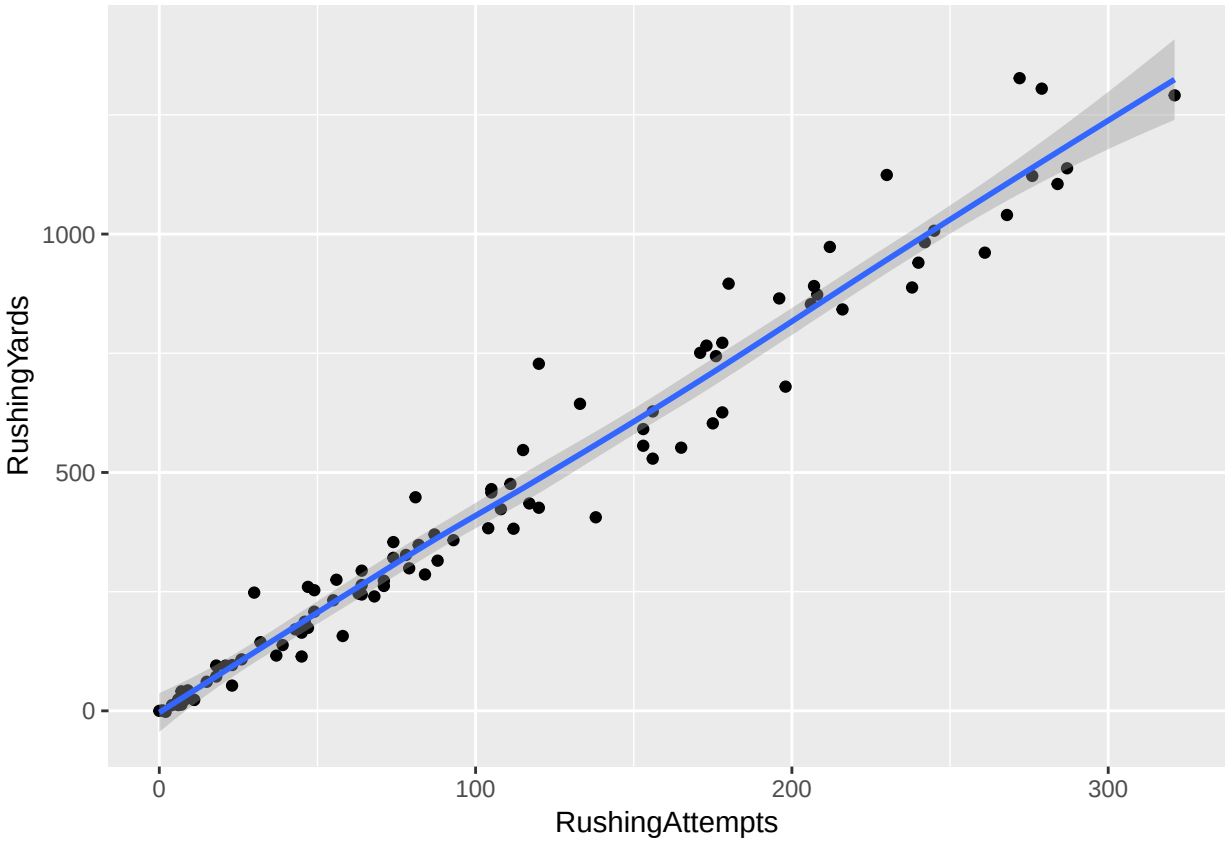
The structure of the dataset was relatively clean, however, I removed any players that had zero in the total fantasy points column. These are players that didn't play enough to generate any material statistics during the season. These outliers are not helpful in trying to predict draft picks for the next season since they didn't earn any points in the 2017 season. Also, any player who scored negative fantasy points for the entire season, which could be an indication of incorrect data, was removed from the dataset.

I also removed a few columns from the dataset that were not relevant to predicting which player would score the most points. Since the entire dataset consists of only running backs, I removed the position column since this was redundant. I also removed the fumbles column, since only the fumbles lost column is needed when determining the amount points scored. A player loses two points for each fumble lost. A fumble that is recovered by the same team does not cause the player to lose any points.

Finally, I joined the average draft position dataset with the main running back dataset, so that there was only one final dataset to work with when diving into plotting and modeling.

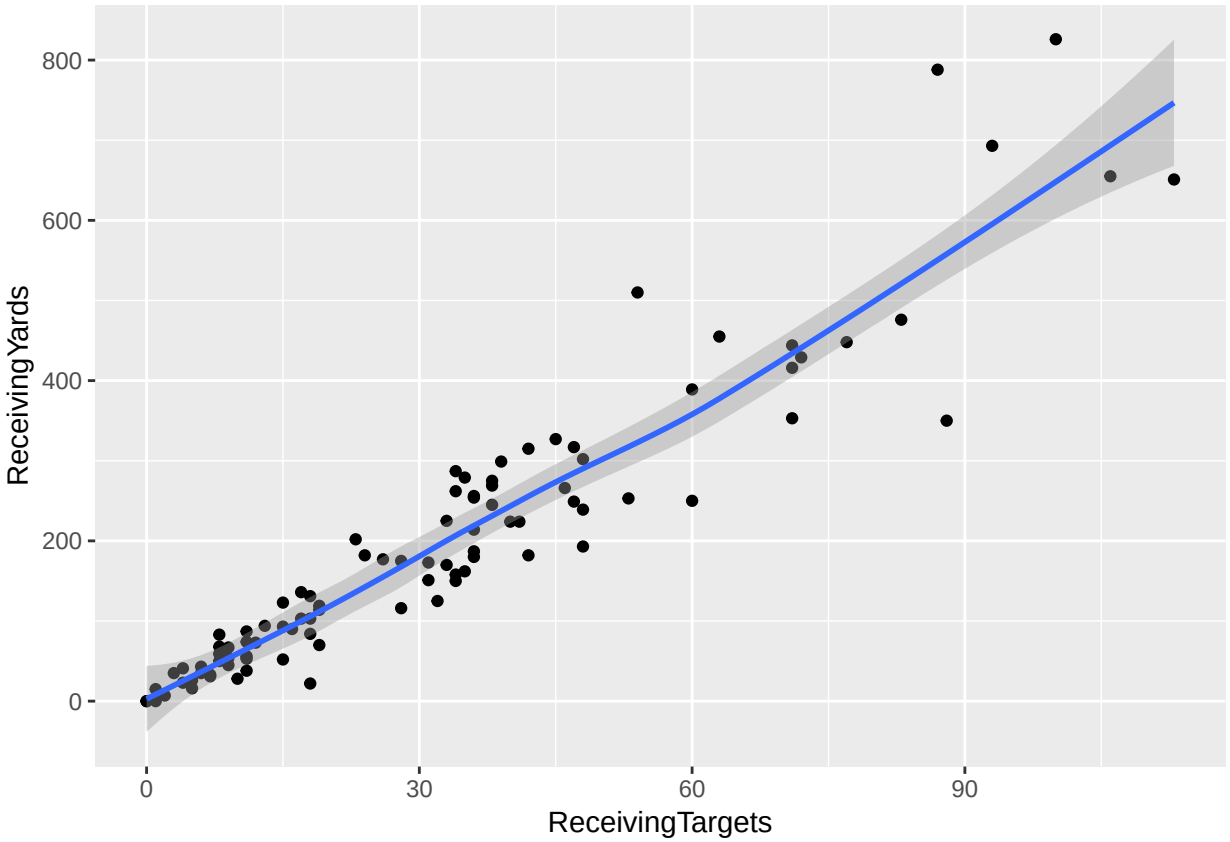
Data Visualization

With the data cleaned up, the next steps involved taking a deeper dive into the data. I decided to create a few different plots to see if anything jumped out right away. The first relationship I looked at was rushing attempts and rushing yards.



As you can see in the plot, there appears to be a linear relationship between the number of rushing attempts a player has over the course of the season and the total number of yards that player rushes for. We will test this theory later, but since more yards equals more points, it initially seems that we would want to focus on players that get the ball more over the course of the season. The team each player is on can have a huge impact on this statistic because it depends on offense strategy that the team implements whether or not they are going to run the ball. As you can see from the plot, in 2017 it seems like Kareem Hunt (on the Kansas City Chiefs), Le'Veon Bell (on the Pittsburgh Steelers), Todd Gurley (on the Los Angeles Rams), and Jordan Howard (on the Chicago Bears) are among the top players with the most rushing attempts.

Another important relationship to consider when trying to predict total fantasy points scored is between receiving targets and receiving yards. In PPR (Points Per Reception), this is even more important since the player receives an extra point for every reception they make. From the graph below, the first thing that jumped out was that this relationship was similar to the rushing attempts and rushing yards relationship in that the more receiving targets a player gets throughout the season the more yards they typically gain. They both have a positive linear relationship.



You can also see that there is a strong correlation between both these relationships, however, there is a slightly stronger one between rushing attempts and rushing yards. This makes sense because typically the more opportunities a player gets to gain yards (both rushing and receiving), the higher chance there is that he will score more points at the end of the season.

```
cor(rb$RushingAttempts, rb$RushingYards, method="spearman")
```

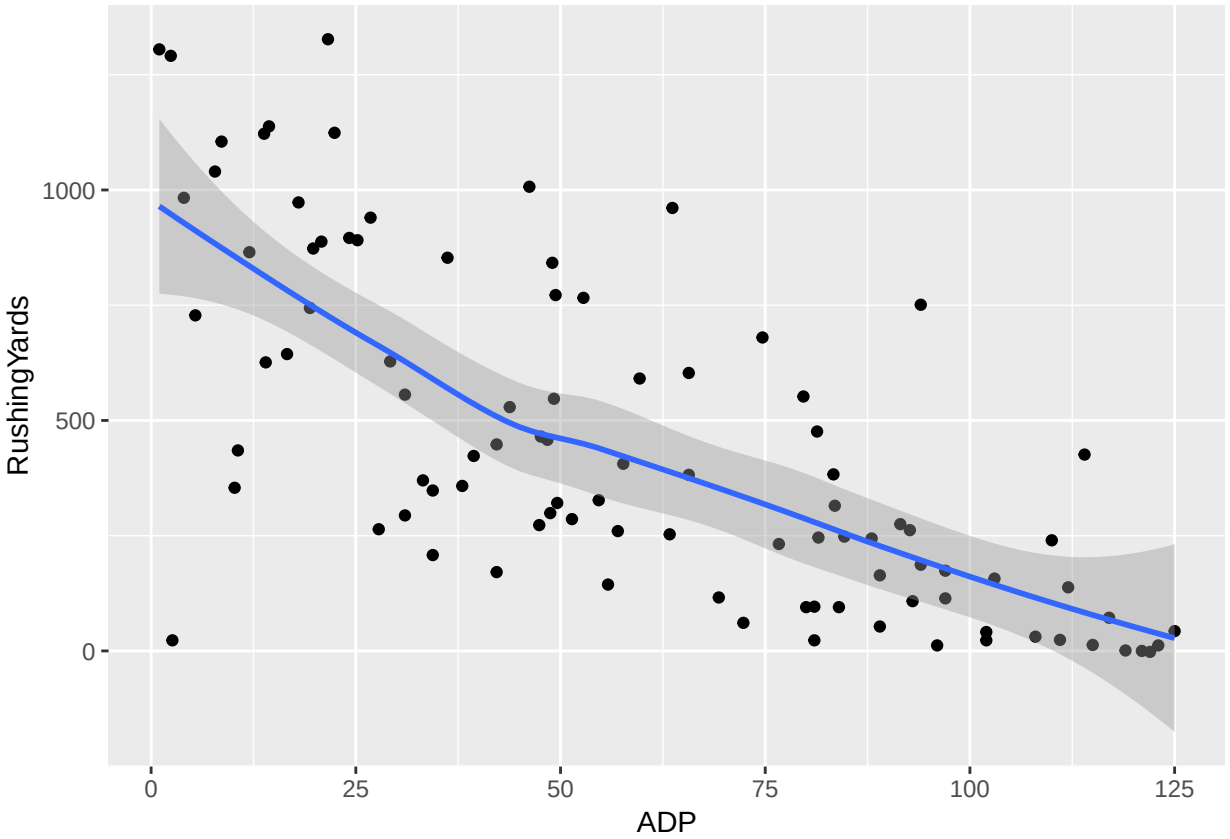
```
## [1] 0.9872207
```

```
cor(rb$ReceivingTargets, rb$ReceivingYards, method="spearman")
```

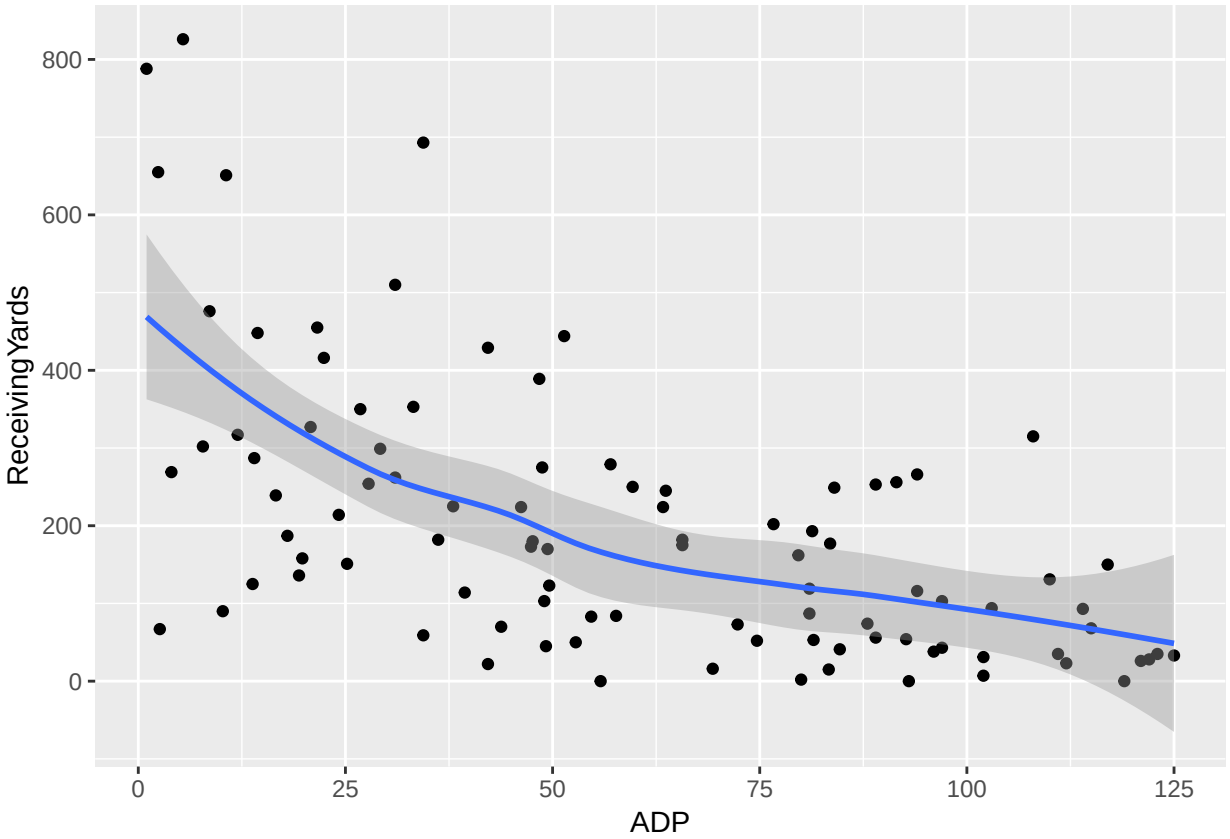
```
## [1] 0.9554589
```

Average Draft Position

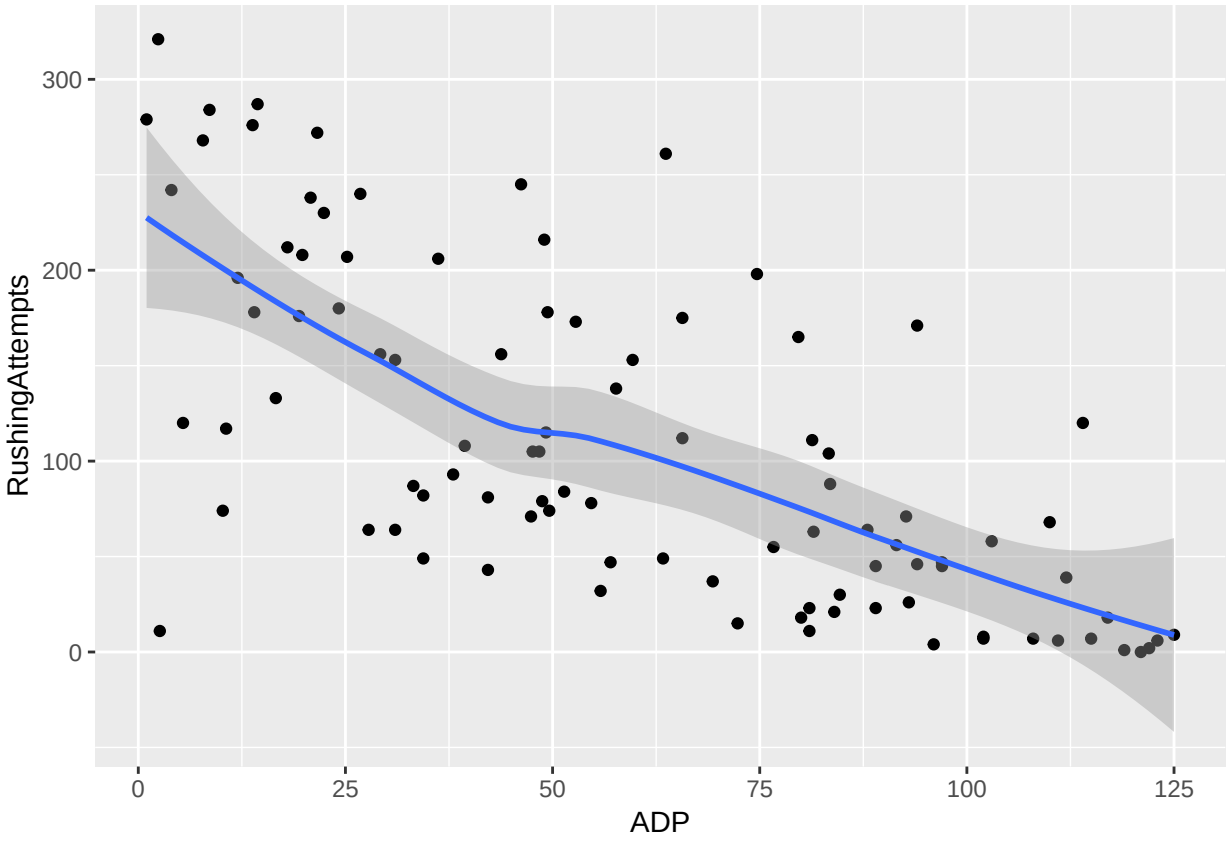
After deciding that these four statstics (rushing yards, receiving yards, rushing attempts, receiving attempts) were the most important in determining how many fantasy points a player will score by the end of the season, I then plotted each of these with the average draft position the following year to see which one of these was the best predictor of future draft position.



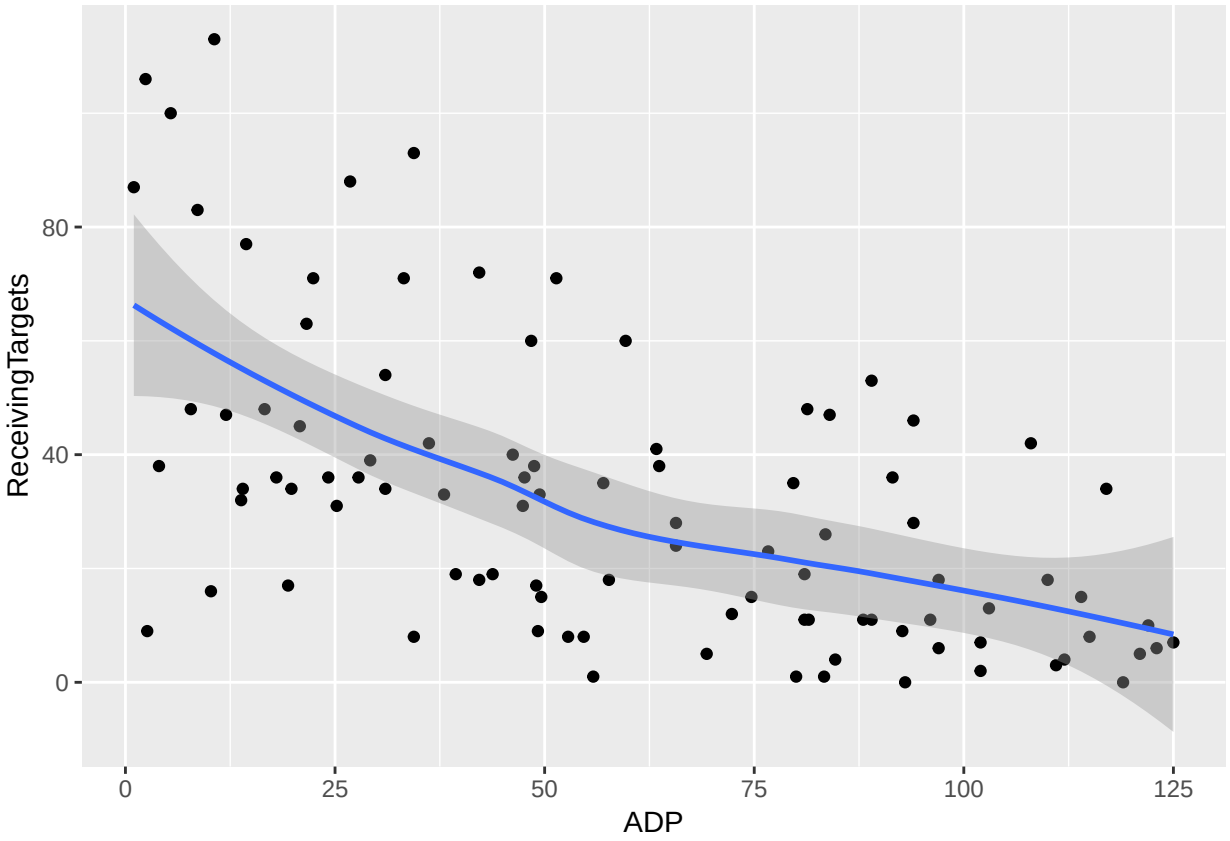
From the rushing yards plot, it appears the majority of the highest total rushing yard players (above the line) were all drafted within the top 25 spots. There were a few outliers, however, which could be due to a couple different scenarios. First of all, if a star player was injured for all of the previous season and is now healthy, then he could still be drafted in the top ten this year, even if he didn't gain any rushing yards the previous season. Also, a highly rated rookie that is expected to perform well over the course of the upcoming season could still be drafted in the first or second round of the fantasy draft. There is a risk in implementing this strategy, however, since there are no pro level statistics available to analyse. A league member who does draft a rookie is expecting them to perform as well at the pro level as they did on the collegiate level. A great example of this scenario in the current year is Saquon Barkley. He was one of the highest drafted running backs this year even though he was a rookie.



The receiving yards plot turned out to be a little more scattered. While the players with the two highest receiving yards were still drafted in the top ten, the rest of the chart was little more difficult to predict. There were several players with less receiving yards than average that were still drafted high. It's important to remember that the scope of this project is only looking at the running back position, so it make sense that total rushing yards would be a little more correlated with draft position. While there are still several running backs who catch a lot out of the backfield, this position is more focused on running the ball.



The next plot I reviewed was the rushing attempts by a player and the average draft position. This represented the total opportunites a player had to run the ball throughout the season. While the total rushing attempts are still important to review, it appears this category is not as statically important as the total rushing yards. Even though there is a strong positive linear relationship between rushing attempts and rushing yards, when it comes to actually predicting the average draft postion, the total rushing yards is more important.



The last plot I decided to look at was the receiving targets. This is the amount of times the running back was targeted in passing game. The receiving targets are similar to the rushing attempts in that each of these plots are more scattered and not as materially important as the total yards a player gains in determing the average draft position the following year.

Linear Regression Models

In order to determine which of these categories was the most staticially significant, I ran several diffent linear models to see which of these was the best predictor for the average draft position. The first predictor I looked at was rushing yards. In this model, the rushing yards are significant indictor of average draft position due to the low std. error and very low probablity that the null hypothesis would be true. In other words, there is very little chance of this result happening just due to random variation and should be considered statitically significant. This model also had a multiple R-squared of .512, which will be need to be compared to the other models.

```
##
## Call:
## lm(formula = ADP ~ RushingYards, data = rb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -86.912 -17.216   1.804  17.747  56.139
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  91.143714   4.040546   22.56   <2e-16 ***
## RushingYards -0.070949   0.007034  -10.09   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.87 on 94 degrees of freedom
## Multiple R-squared:  0.5198, Adjusted R-squared:  0.5147
## F-statistic: 101.8 on 1 and 94 DF,  p-value: < 2.2e-16
```

The next model I reviewed was receiving yards. This model also had a low Pr value, but the multiple R-squared was also significantly lower at .3475. Both these models only looked at only one categorie when predicting average draft position. I decided to keep running models while adding categories to see if there was a better fit. During this process I took into account the idea of overfiiting and tried to keep the focus on only a few categories.

```
##
## Call:
## lm(formula = ADP ~ ReceivingYards, data = rb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -72.174 -20.894   4.093  20.677  62.586
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    82.70620    4.42460  18.692  < 2e-16 ***
## ReceivingYards -0.11839    0.01673  -7.075 2.64e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.99 on 94 degrees of freedom
## Multiple R-squared:  0.3475, Adjusted R-squared:  0.3405
## F-statistic: 50.05 on 1 and 94 DF,  p-value: 2.642e-10
```

After running serveral other regression models, I finally ended up with rushing yards and receiving yards together to predict average draft position.

```
##
## Call:
## lm(formula = ADP ~ RushingYards + ReceivingYards, data = rb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -88.339 -16.292   1.744  17.299  47.157
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    96.259226   3.995527  24.092  < 2e-16 ***
## RushingYards   -0.055906   0.007629  -7.328 8.36e-11 ***
## ReceivingYards -0.060221   0.015570  -3.868 0.000204 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.21 on 93 degrees of freedom
## Multiple R-squared:  0.5863, Adjusted R-squared:  0.5774
## F-statistic: 65.91 on 2 and 93 DF,  p-value: < 2.2e-16
```

As you can see from the below anova review, the model with rushing yards and receiving yards together

turned out to be the best fit. There was the biggest reduction in the residual sum of squares (RSS) between the first and second model. The Pr value was also very low and the most significantly signficant when compared to all four models.

```
## Analysis of Variance Table
##
## Model 1: ADP ~ RushingYards
## Model 2: ADP ~ RushingYards + ReceivingYards
## Model 3: ADP ~ RushingYards + ReceivingYards + RushingAttempts
## Model 4: ADP ~ RushingYards + ReceivingYards + RushingAttempts + ReceivingTargets
##   Res.Df   RSS Df Sum of Sq       F    Pr(>F)
## 1     94 58139
## 2     93 50083  1    8056.2 14.7123 0.0002306 ***
## 3     92 49860  1     222.9  0.4071 0.5250301
## 4     91 49830  1      29.8  0.0545 0.8159674
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ideas for Further Research

There is a lot of room to expand the scope of this project in further research. This project only analyzed one of the many positions in football. I think it would be a great idea to provide the same type of analysis for each position and then tie it all together to get the full picture of a complete fantasy team. Each position could be compared to each other to see which one is the most valuable and which positions to prioritze when drafting. With these analytics, you could help prove the common idea that drafting a kicker in the third round is a terrible decision.

Conclusion

While there will always be several factors to decide on in real time during the fantasy draft, it would be a good idea to develop a solid game plan of the top running backs you believe have the best chance to succeed before the acual draft. While making this list of running backs, there should be a strong focus on his rushing and receiving yards from the previous season. These categories from the previous year could be a strong indictor of future success and should be factored into your ranking. The model with rushing and receiving yards could also be applied on future seasons. Hopefully with this knowledge, fantsy football league memebers will be better prepared for their future drafts.