Exploratory Data Analysis

of

The Distinction of Overachieving NFL Players Relative to Their Position Group Through Machine Learning

By

Anthony Grieco and Zachary Phillips

Source: <https://www.pro-football-reference.com/>

Submission Date: 04/30/2024

Foundations of Data Science and Analytics - CS345

Professors: Dr. Adam R. Albina, Dr. Stephen Shea

Introduction

With the popularity and societal impact of American Football flourishing under the direction of the National Football League (NFL), it has never been more important for its thirty-two member teams to carefully analyze how they can better perform as a team. This starts with both the NFL Draft and Free Agency, where teams look to fill holes on their respective rosters and build the most-competitive team that they can with limited monetary resources. As such, part of the scouting process involves identifying which position groups, and more specifically players, are the most impactful on the field. With the emergence of Machine Learning, it has never been easier to identify which members of pre-established position groups perform the best in a league that has its quarterbacks predominantly throw the ball as opposed to simply handing it off to a teammate. We intend to create a model that identifies position-based outliers and misclassifications in clusters created with Machine Learning in order to help teams identify which players are the best at their respective position groups from a receiving perspective. We also hope to identify clusters of new and emerging hybrid position groups so that we can better encompass the changing nature of both the game and the NFL. Our hope is that teams would be able to use our model to help them better identify which players they should pursue in both the NFL Draft and Free Agency in an otherwise pass-happy league now and in the future.

The data that we used in this exercise was obtained from Pro Football Reference, a well-respected football-related data collection and analytics company, and consisted of individual player statistics from 2014 to 2023. The specific dataset that we used in this experiment consisted of ten individual data sets each containing the same types of player data but from their own respective league years ranging from 2014 to 2023. We decided to aggregate the data from each of these individual data sets so that we could compare whether a player’s performance was consistent across multiple consecutive years or if they instead had a single good year. Players that consistently perform better over longer periods of time are those that NFL teams covet and are more likely to pursue in Free Agency and we wanted to have a way to measure that statistically through Machine Learning.

The analysis questions that we attempt to answer are as follows:

1. Through unsupervised Machine Learning, we hope to determine the accuracy of our model for determining whether a player in the NFL over the last ten years (2014 - 2023) was a Wide Receiver, Running Back, Fullback, or a Tight End. Our model will focus exclusively on a player’s receiving stats including the total number of receptions, total receiving yards, average yards per reception, and the longest reception that each player had in the NFL in that same time frame. Any players with excellent receiving stats that the model misclassifies as another position group are those that teams ought to pursue as they have consistently outperformed expectations for their respective position group.
2. Through unsupervised Machine Learning, we hope to identify clusters in the data that indicate which players, regardless of their listed position, are most similar to one another from a receiving standpoint.
3. Through unsupervised Machine Learning, we hope to be able to more accurately identify different sub and hybrid positions of all receiving players regardless of their current official NFL position. For example, based on the clusters we derive in Question 2 above, we hope to identify whether players in a certain cluster more closely align with being a boundary or a slot Wide Receiver as opposed to simply being known as a “Wide Receiver” because there is currently no official distinction between the different types of players who play the same position.

Basic Data Cleaning and Preprocessing

The biggest issue that we ran into with the final data set that we ended up using in our experiment was that we couldn’t find a single complete set of data that contained NFL receiving stats for every player that has played in the league over the last 10 years. This meant that we had to manually take each individual year’s player data and then aggregate them into a single master data set. We then placed a filter on the data set so that we would only have data on players who had at least twenty targets (see line 10 of the attached R script). This allowed us to weed out the players at positions who most consistently saw the ball thrown to them. We then checked our data for either missing or incomplete values, but no such data existed after the filter was applied (see line 14 of the attached R script). We also manually removed any extra characters under each record’s name variable, such as “\*” and “+”, that were used to indicate different honors that a particular player received at the end of that season including “Pro Bowl” and “All Pro” respectively. That way we could achieve consistency in the data as a single player’s progress could be tracked over at most ten years without the possibility of having the same player be represented multiple times in the same model simply because they received different honors in different years. For example, if a player received “All Pro” honors their first year, and was given a “+” sign next to his name that year, but then didn’t receive that same honor in the next, and therefore didn’t have the “+” sign next to his name, the original data set would have classified these two entries as two separate players even though both entries were actually referring to the same player. This predicament is no longer the case in our new combined data set. Additionally, we set multiple columns to be factors including year, position, and team in order to more accurately represent the types of data we were working with for our analysis (see line 33 of the attached R script). Finally, we cloned our original data set and standardized every numerical variable so that we could see if there were any differences in our findings after every plot we ran, but we ultimately found no significant changes between the raw and standardized data sets in any test that we ran (see line 38 of the attached R script).

Univariate Exploration

As a means to more thoroughly explore our new data set, we created a series of histograms and boxplots in order to identify potential trends in the data. We started by mapping the total number of players we had by position in a bar chart and found that our data set contained at least twice as many Wide Receivers as any other individual position group across the total ten years our data was derived from (see line 45 of the attached R script Appendix A). This makes sense because our data set had already been filtered so that only players with at least twenty receptions would remain in our data in the first place. We then took a look at targets, or the number of times a receiving player was thrown to by the quarterback, and found that the data was positively skewed. This indicates that most players received around the same number of catches, and it was only either the best or most utilized receiving players, who were relatively few, who saw more targets than every other receiving player (see line 50 of the attached R script and Appendix B). Similar trends could be found in the total number of receptions made per player per season (see line 59 of the attached R script and Appendix C), the total number of yards per player per season (see line 68 of the attached R script and Appendix D), and the average number of receptions made a game per player (see line 95 of the attached R script and Appendix G). Although the average yards per target (see line 77 of the attached R script and Appendix E) and per reception (see line 86 of the attached R script and Appendix F) respectively had more normal distributions, the principle described above largely remained the same as most players had comparable average yards per target and reception to one another with few players exceeding these benchmarks. This was to be expected as there are different tiers of receiving players in the NFL and it is on these very same receiving players who stand out from their peers that we intend to run our unsupervised machine learning models against.

Bivariate Exploration

We continued to explore our data set by looking at the relationships between our data through a series of scatterplots. We started by looking at how the number of targets thrown to receiving players compared to the total number of yards that those targets were for (see line 106 of the attached R script and Appendix H) and found that there was a highly positive correlation between the two. This was exactly what we expected because intuitively the more often a player was targeted, the more yards the ball would have traveled when it was thrown. This was supported by a covariance of 11282.48 and a correlation of 94.86%. We had similarly expected findings for the relationships between targets and receptions with a correlation of 96.62% (see line 114 of the attached R script and Appendix I), receptions and yards with a correlation of 92.29% (see line 122 of the attached R script and Appendix J), receptions and touchdowns with a correlation of 69.31% (see line 130 of the attached R script and Appendix K). Finally, we ran a histogram in order to see the total number of players who played at each particular position in the NFL from 2014 through 2023, which allowed us to get a better understanding of the direction that the league is currently heading in from a receiving standpoint (see line 138 of the attached R script and Appendix L). Interestingly, the number of Running Backs who had at least twenty receptions in a given year has steadily decreased over the last ten years whereas the number of Wide Receivers in the league with at least twenty receptions a year has generally increased more rapidly in that same time period. The number of receiving tight ends in the league with at least twenty receptions have also generally increased which, even without an outside understanding of the current Running Back market, indicates that teams are slowly moving away from Running Backs and are instead turning towards Wide Receivers and Tight Ends for their future in the receiving game. One final item we noticed in this chart is that there appears to have only been one instance where a Quarterback had twenty or more receptions, which is an extraordinarily rare feat for a QB as indicated by this happening only during the recent 2023 season over the course of the last ten years.

Multivariate Exploration

Our final exploration consisted exclusively of scatterplots, and each sought to visually draw out the specific differences between each position group in our data. As a way to further experiment with trends we had already identified in our data, we decided to run some of the same tests we ran before during our Bivariate Exploration but did so in order to see how different positions groups generally performed relative to one another. We started by reviewing the number of yards per target and then adjusted our scatterplot so that it would indicate whether a particular data point belonged to a Fullback, Quarterback, Running Back, Tight End, or Wide Receiver (see line 145 of the attached R script and Appendix M). Doing so revealed that although the data still followed the path of a very strong positive correlation, Wide Receivers and Tight Ends could be found all over that path board whereas Running Backs and Fullbacks were generally found more towards the lower end of this path. The next relationship we looked at with position in mind was every target thrown vs receptions caught, where we found that Running Backs generally seemed to make more receptions for every target they received followed by Tight Ends being the second-most accurate catchers and finally with Wide Receivers having the fewest receptions for every target thrown their way (see line 150 of the attached R script and Appendix N). This can most likely be attributed to how Quarterbacks generally pass the ball shorter distances to get the ball to Running Backs whereas Quarterbacks more frequently have to launch the ball much further in order to get it to the hands of their Wide Receivers with Tight Ends generally having to make catches that fall between the above two position groups. Generally the shorter the pass made, the more accurately it will get to its intended receiver. This would explain why Running Backs, who generally specialize in short-yardage situations, have a higher catch percentage than Wide Receivers who thrive in deep coverage do. This exact trend is explained in our yards per reception by position scatterplot, where Running Backs generally had receptions for far fewer yards than Wide Receivers had (see line 155 of the attached R script and Appendix O). Finally, our touchdowns per reception scatterplot had a much higher concentration of Tight Ends and in particular Wide Receivers who not only made a reception that resulted in a touchdown relative to Running Backs, but we also had far more instances where those same Tight Ends and Wide Receivers made multiple receptions in a single season that went for touchdowns compared to that of Running Backs (see line 160 of the attached R script and Appendix P). Of the limited number of Running Backs who made receptions that led to touchdowns, most did not have any more than five passing touchdowns whereas most Wide Receivers did not have any more than ten passing touchdowns.

Conclusion of the EDA

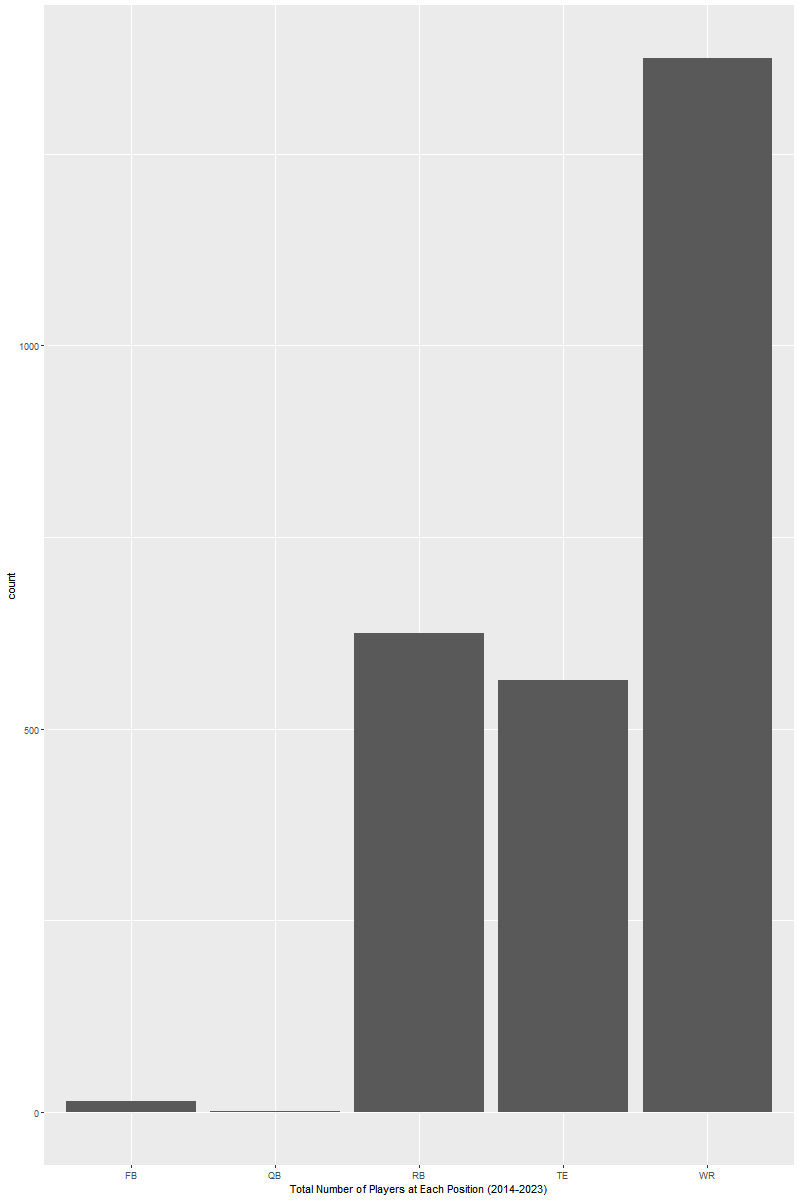
The goal of our Exploratory Data Analysis was to identify which characteristics generally distinguish the five main position groups from a receiving standpoint in the NFL from one another. Based on our exploration using a series of bar charts, histograms, box plots and scatterplots, we have identified that total targets, yards, receptions, average yards per reception, catch percentage, and touchdowns are all potentially significant contributors for determining whether a player is a Wide Receiver, Running Back, Tight End, Fullback, or a Quarterback. Much of what we found in our data set through exploration was what we expected to see, such as Wide Receivers in general having far more receiving yards than Running Backs over the last ten years for example, but the models we ran also helped us visualize other parts of the game that aren’t so apparent but intuitively make sense. One example of this is how Running Backs generally have a greater catch percentage than Wide Receivers do because the receptions that they do make are generally in short-yardage positions where the Quarterback can simply toss the ball to them a couple yards away from the line of scrimmage rather than being forced to launch it to a Wide Receiver far down the field where a completed pass becomes far less likely. All of the trends we have identified during this exploration phase will help us answer our analysis questions above in our final analysis phase.

Statement of Analysis

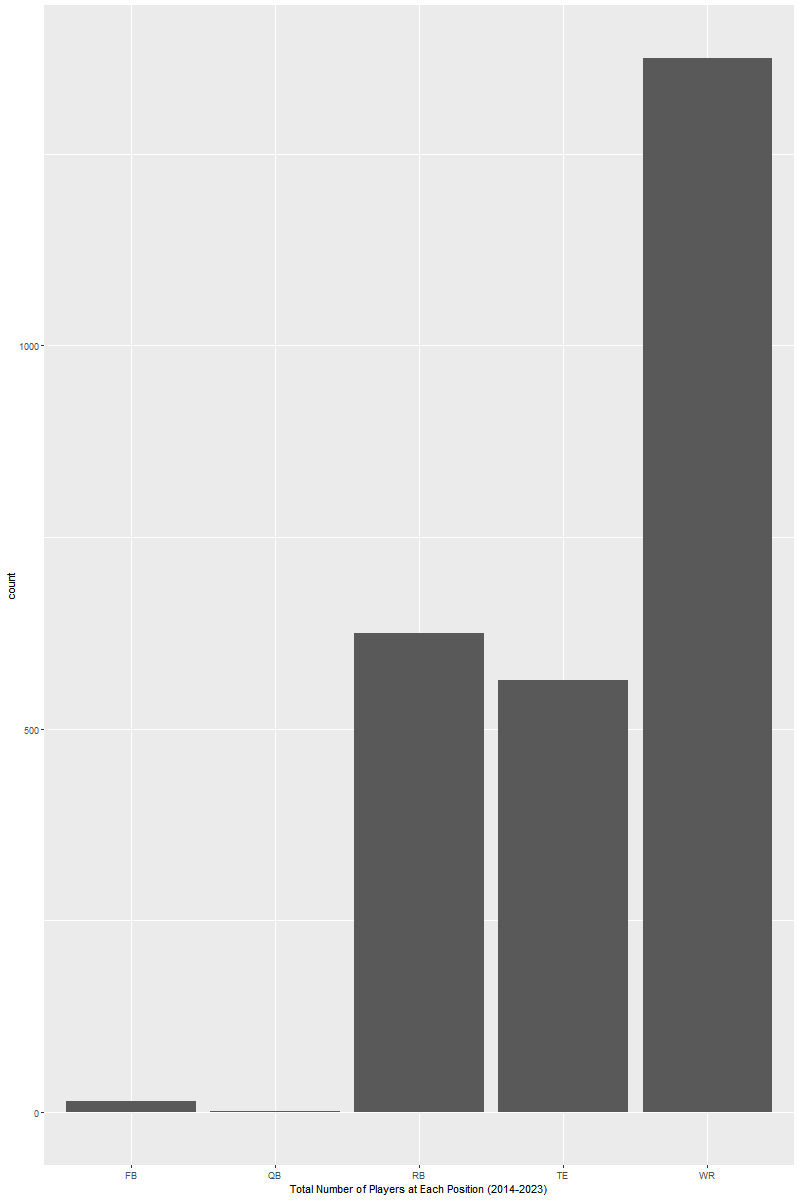
Using the factors that we have identified above, we intend to build unsupervised Machine Learning models that will allow us to answer our above analysis questions. These will allow us and actual NFL teams to not only determine what position groups in the future might look like, but also help current teams identify high-impact players in the receiving game relative to peers in their same position group. The idea is that players of one position group who are automatically and mistakenly clustered alongside elite players of another position group by our model will be identified as more desirable players for teams to target either in the NFL Draft or Free Agency. This would be the case because these players are outperforming their position group peers and therefore have the potential to have higher upside for the team seeking their services than their peers. This is especially true when this occurs on a consistent basis across multiple seasons as a single elite season can be the result of many different factors. If a player consistently outperforms their peers from a receiving context, then that is a player that every team in the NFL should want to have on their roster in order to have the best chance of winning the Super Bowl.

Appendix A: Total Number of Players at Each Position (Univariate)

Raw Data:

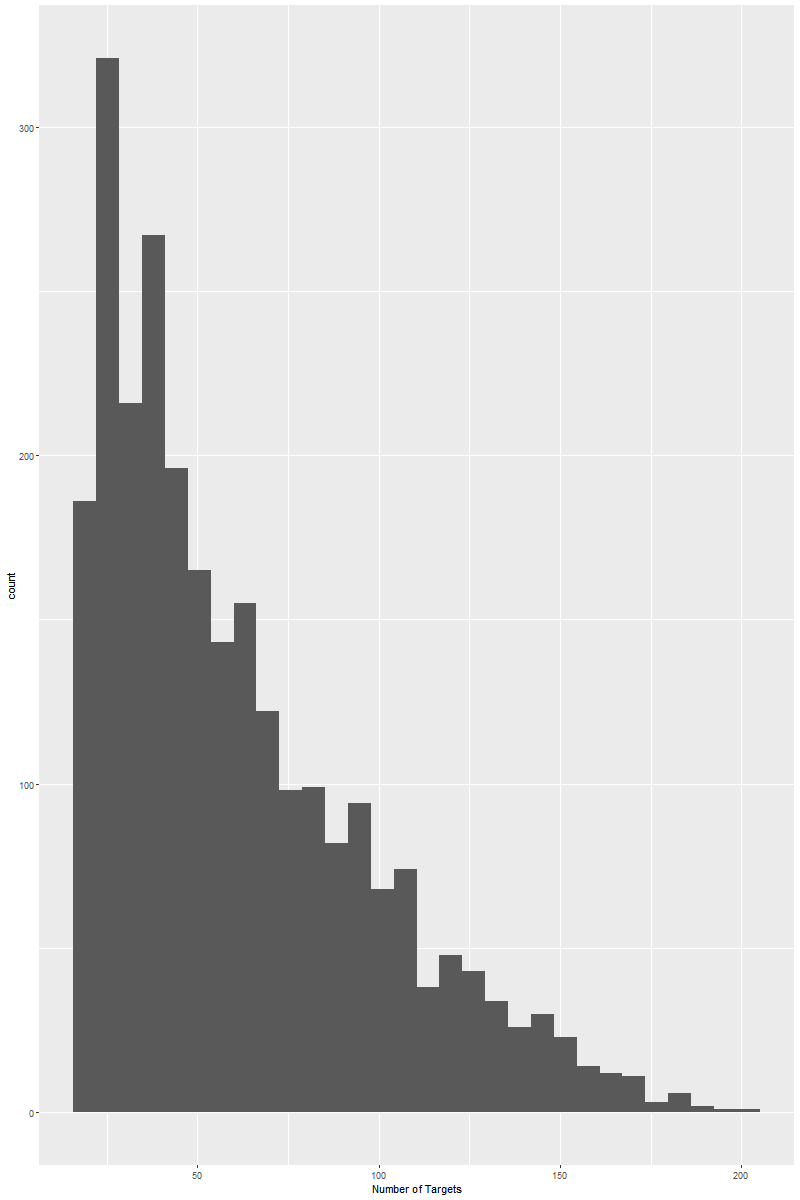


Standardized Data:



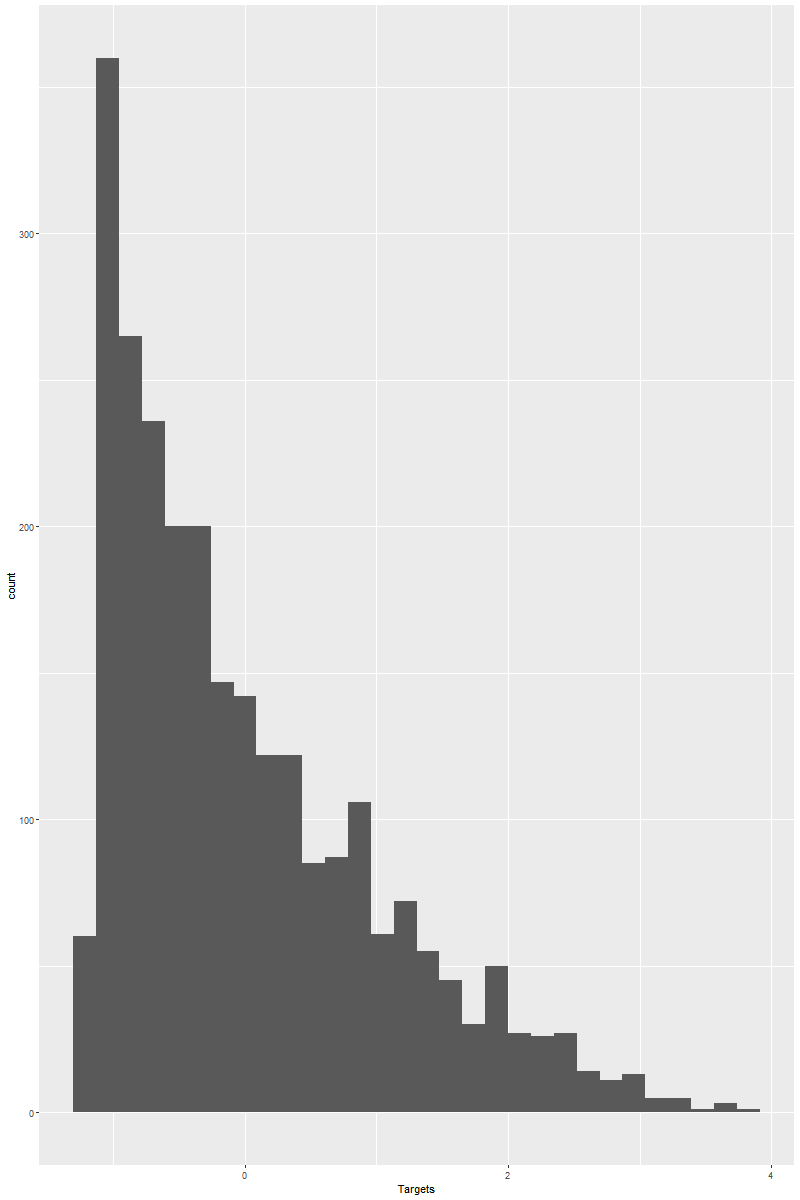
Appendix B: Number of Targets Per Player (Univariate)

Raw Data:





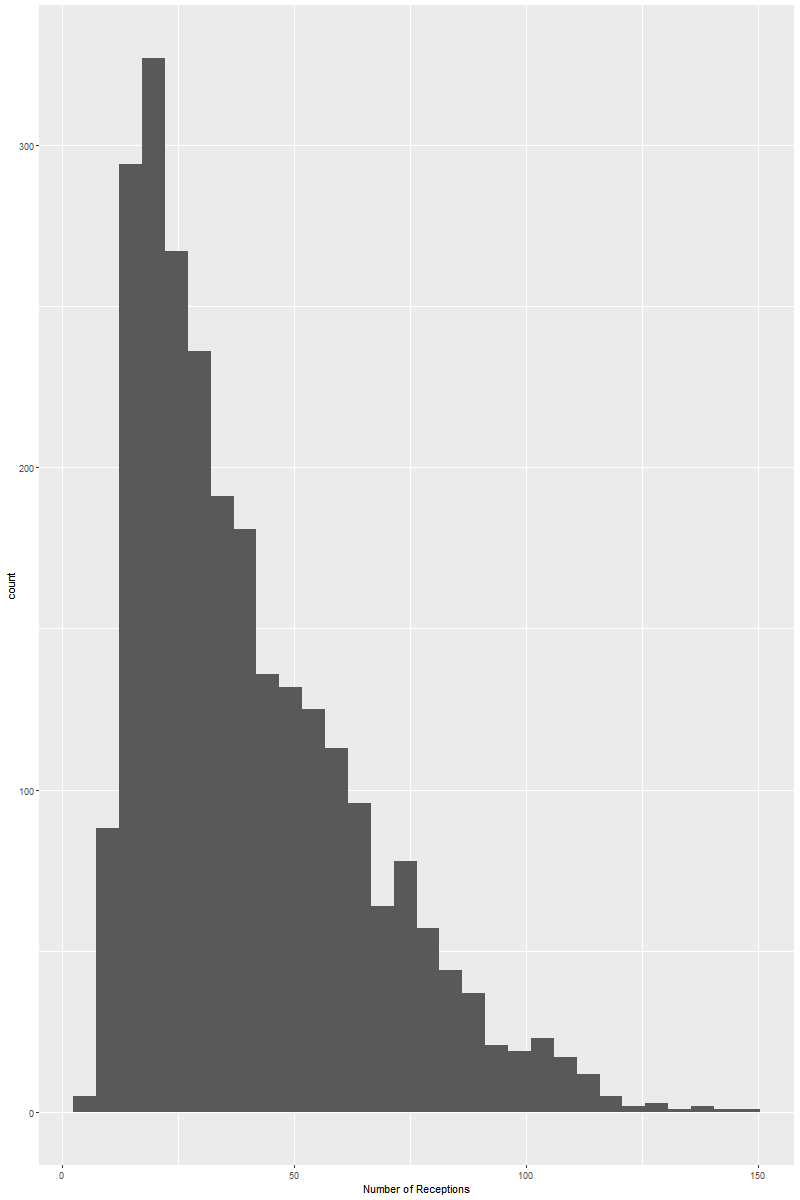
Standardized Data:

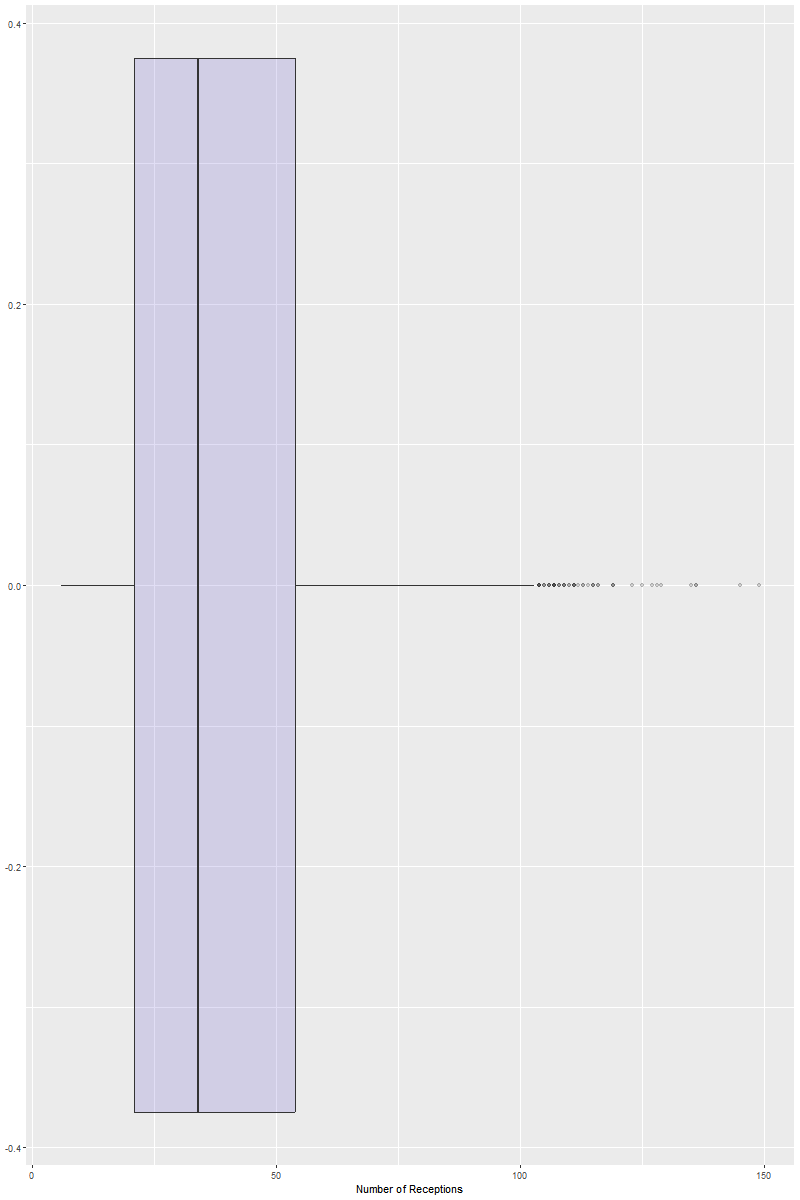




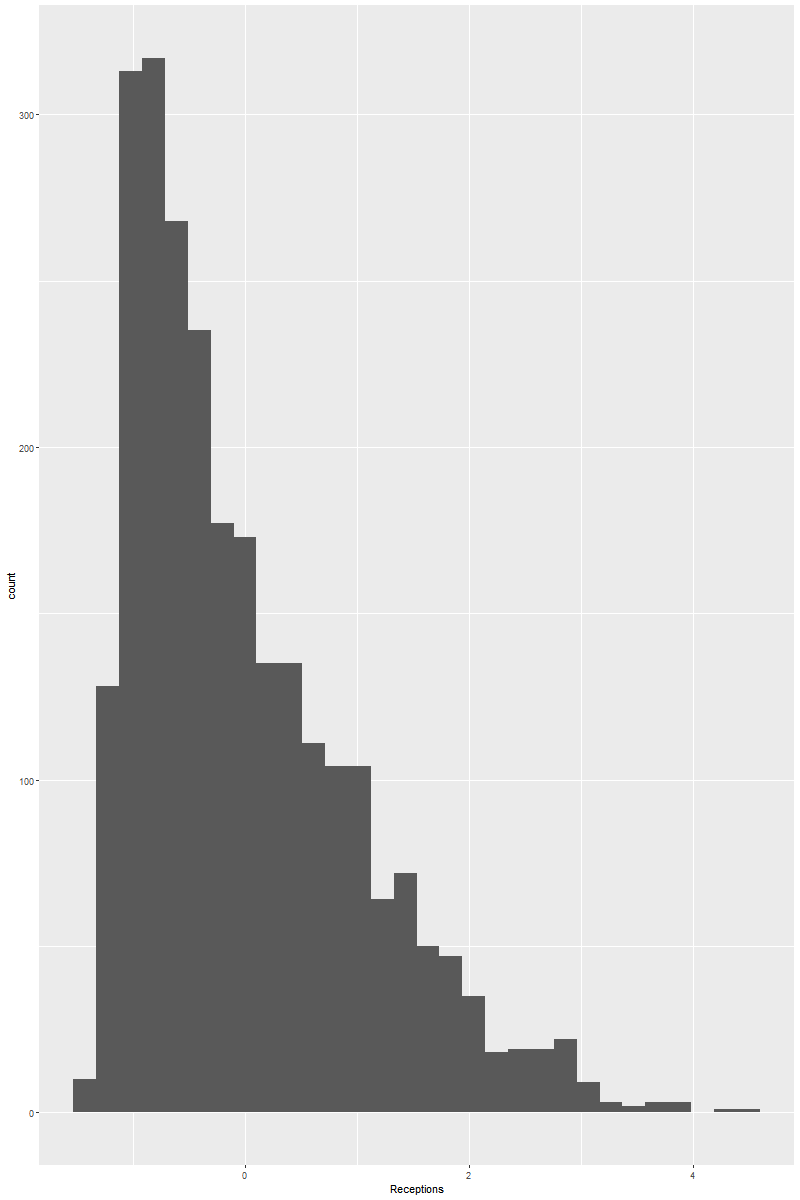
Appendix C: Number of Receptions Per Player (Univariate)

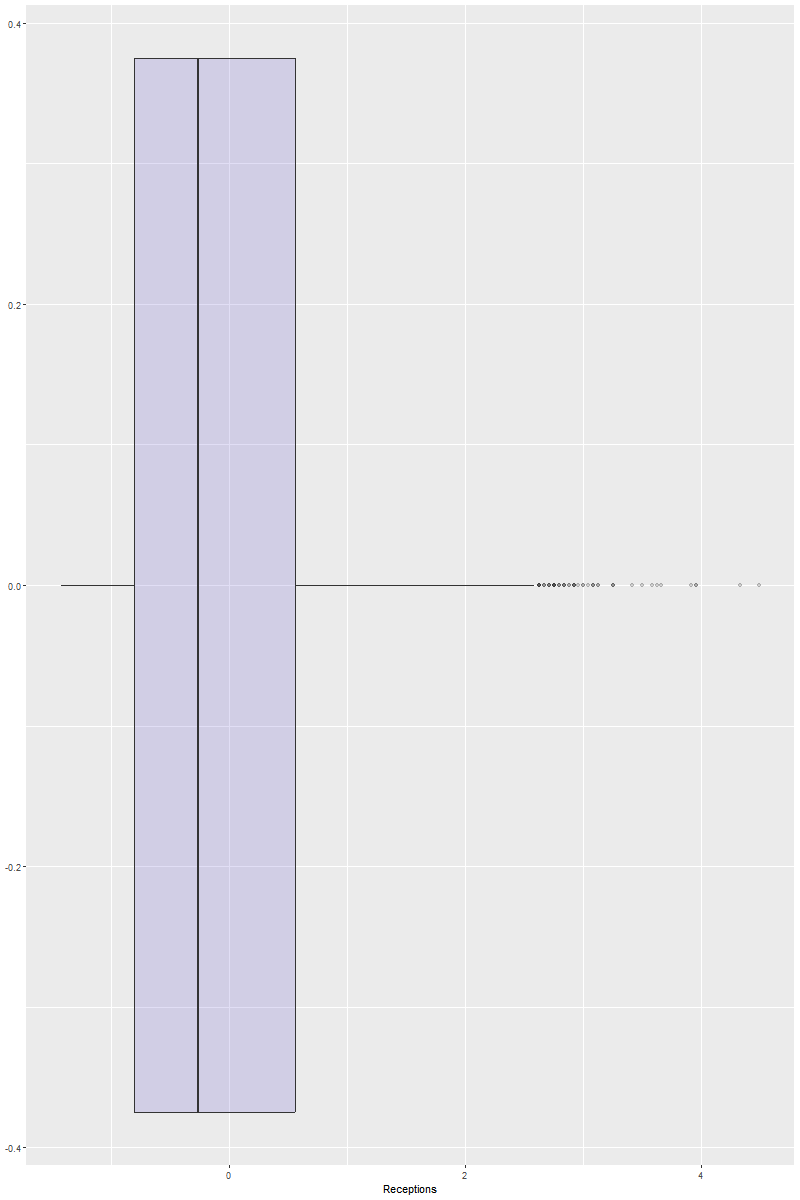
Raw Data:





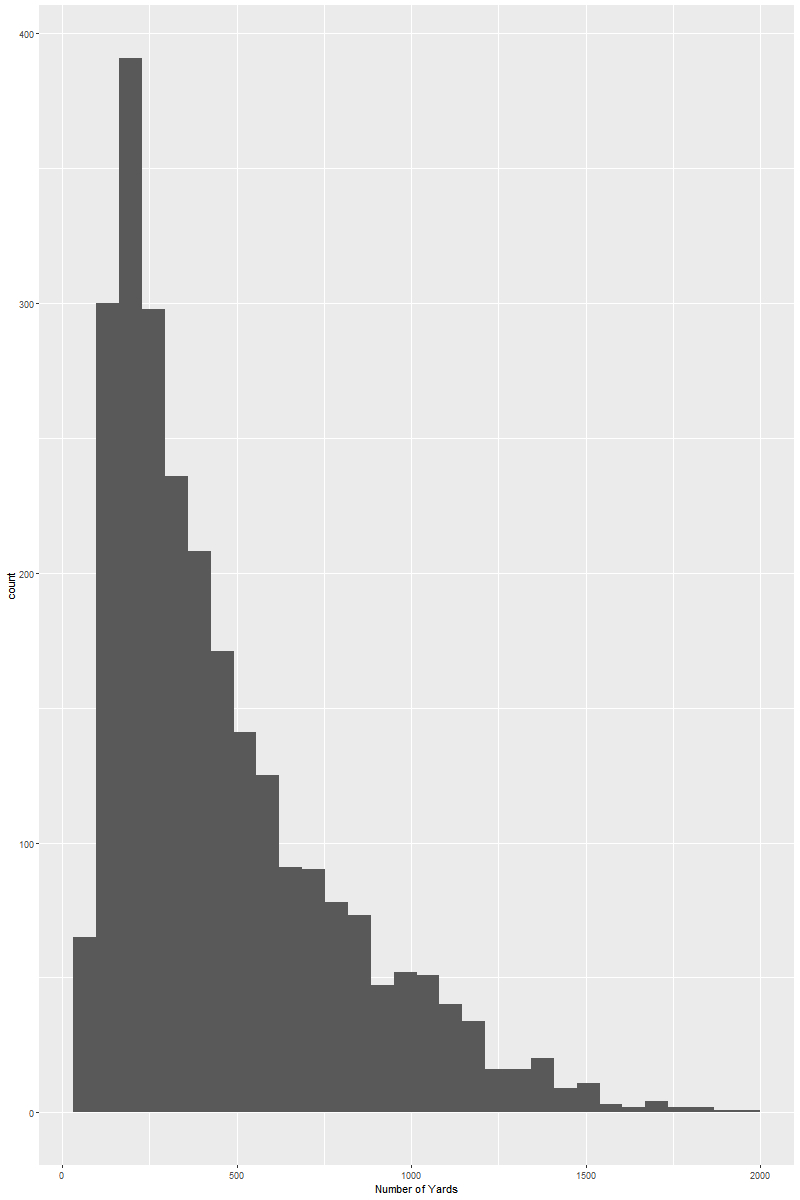
Standardized Data:

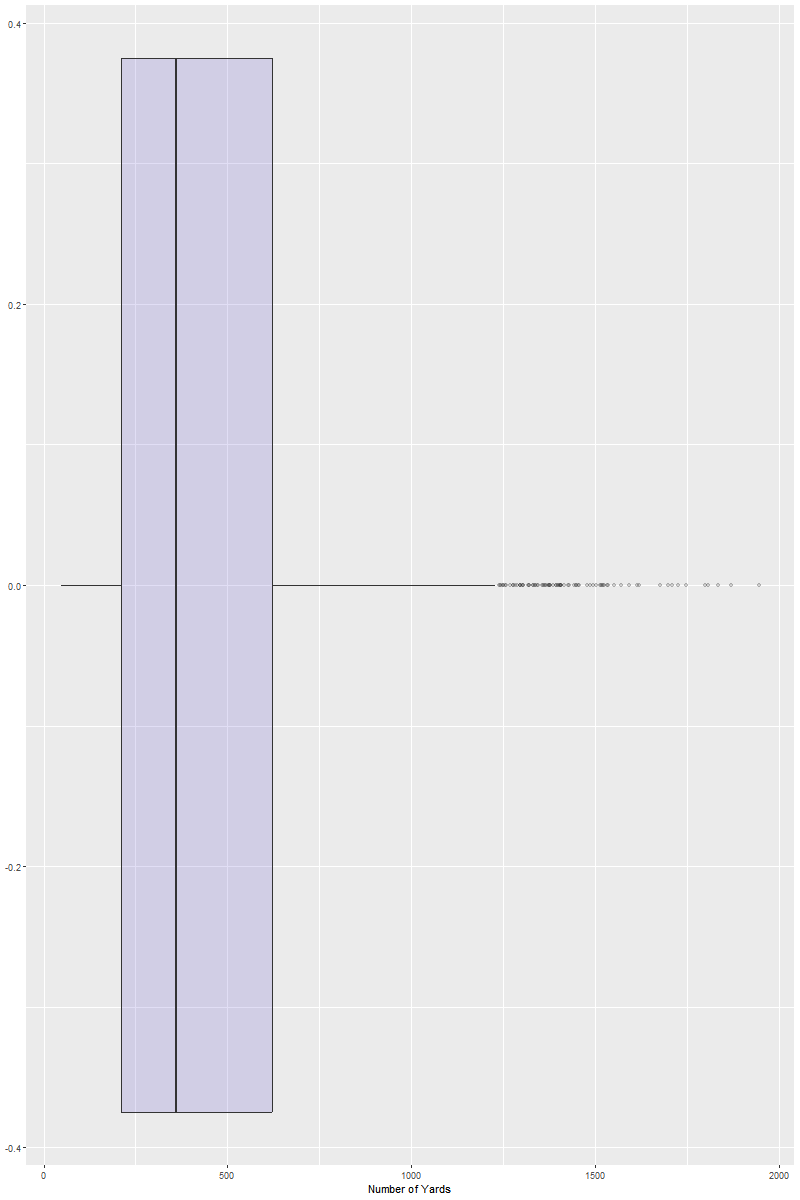




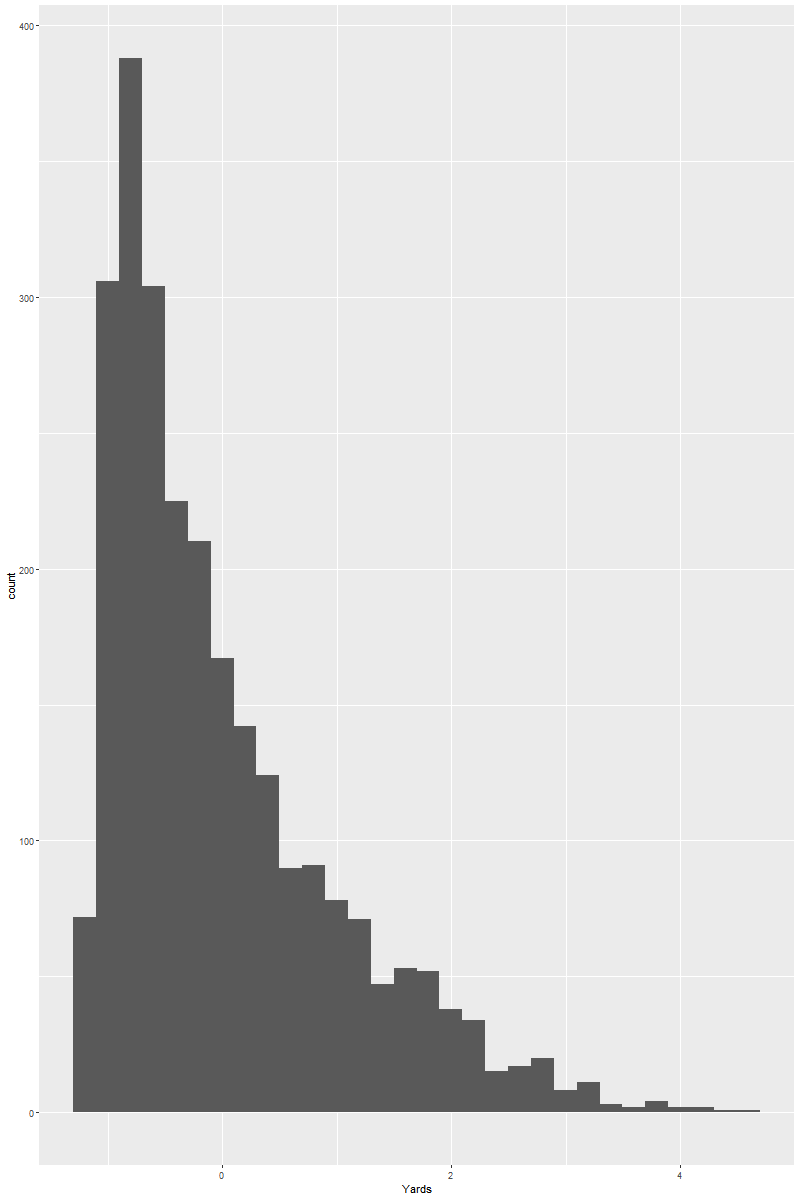
Appendix D: Number of Yards Per Player (Univariate)

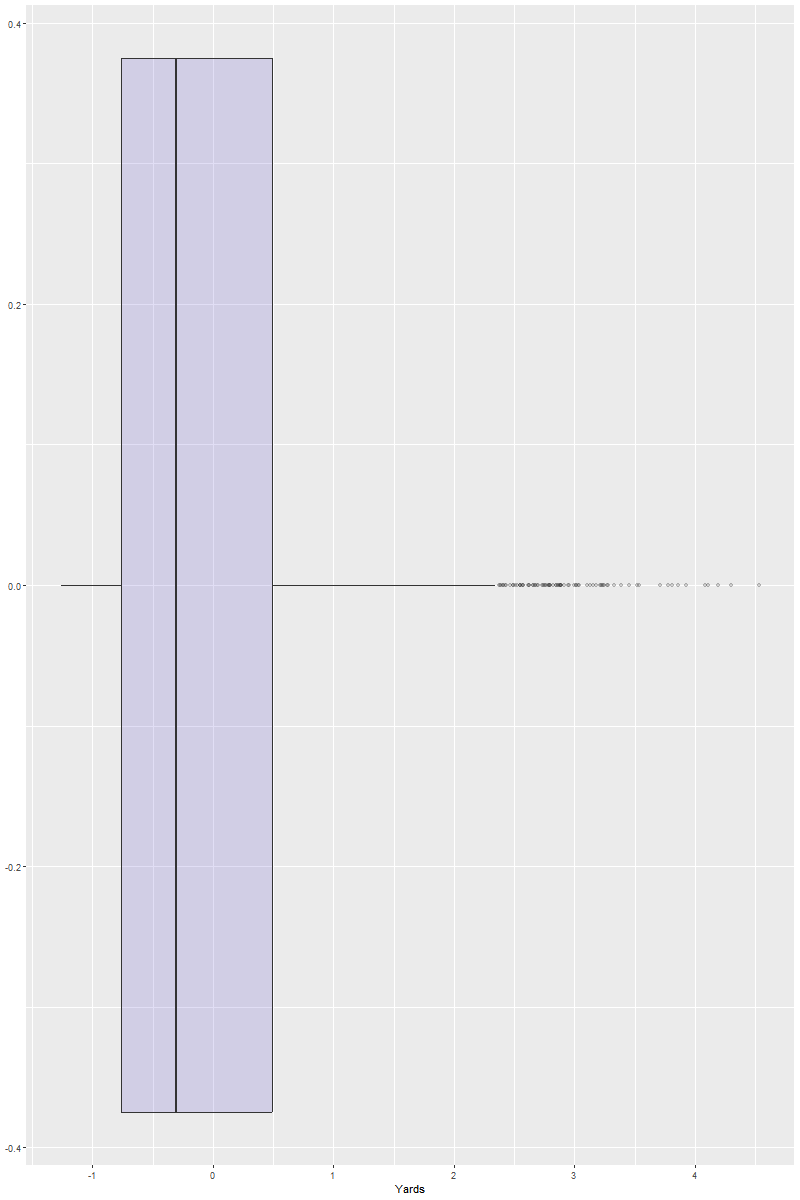
Raw Data:





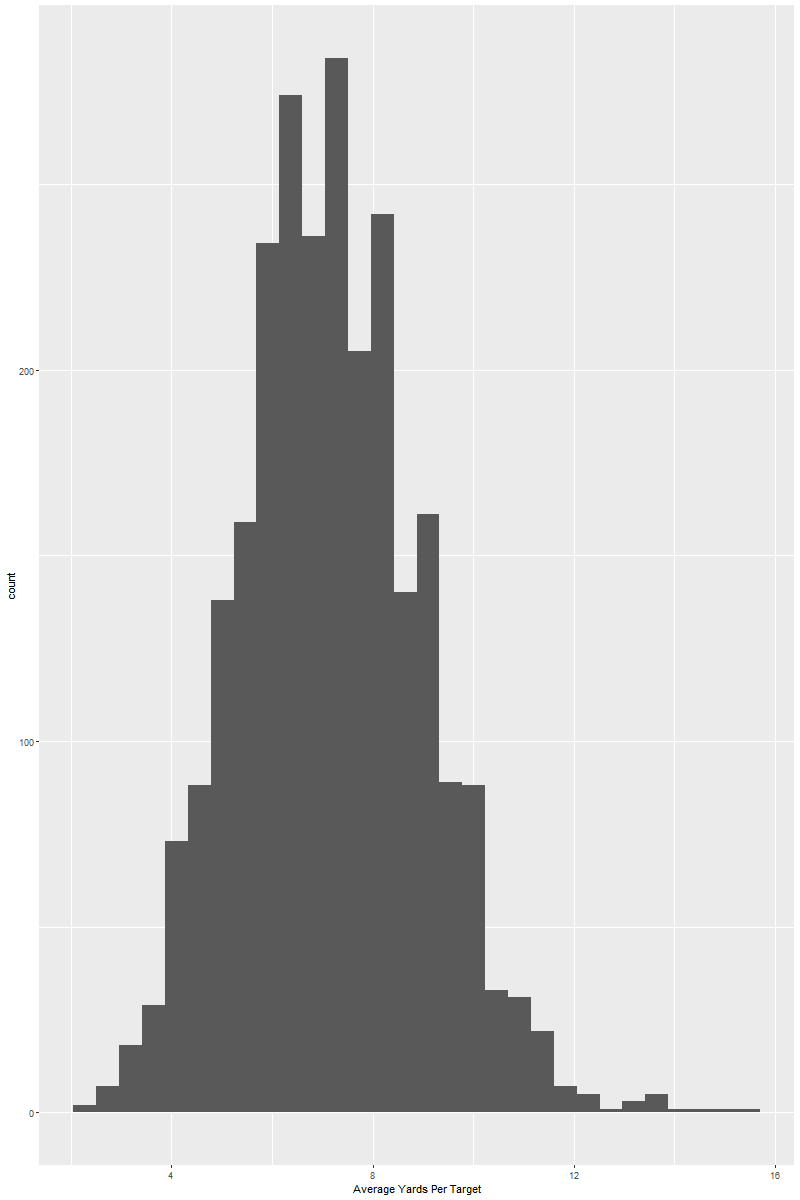
Standardized Data:

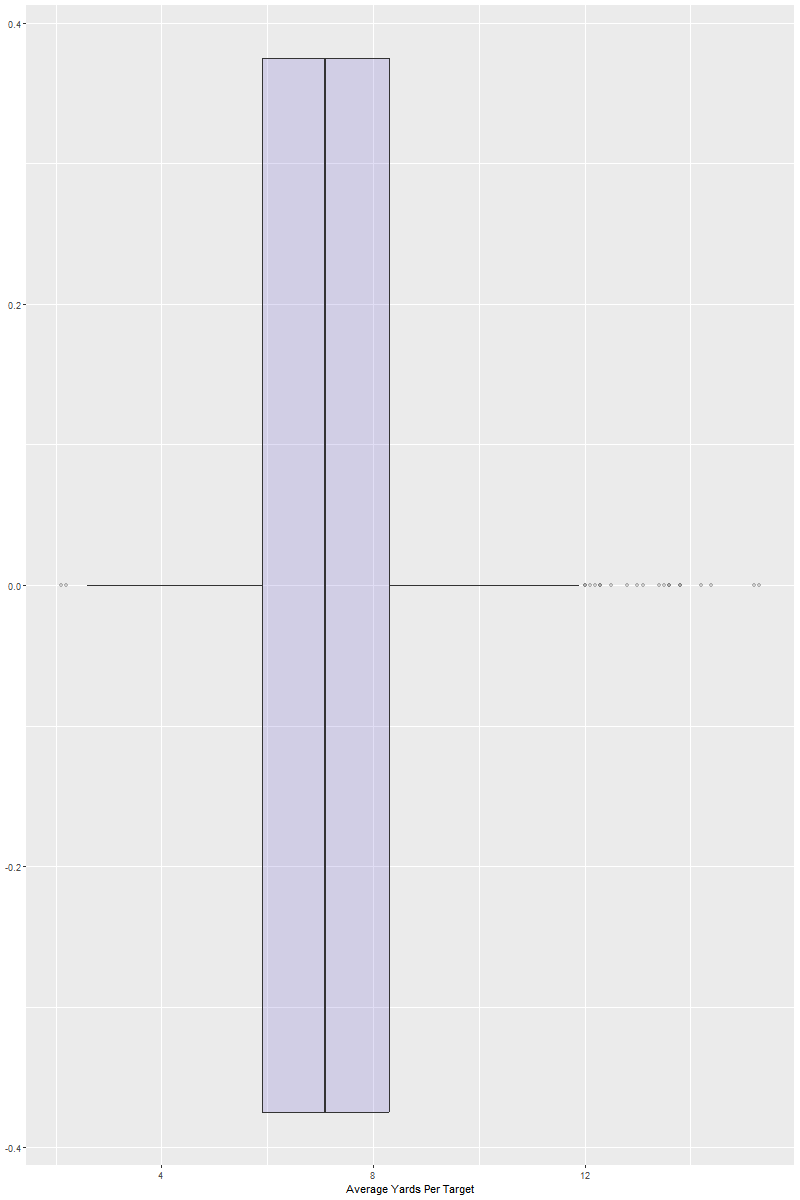




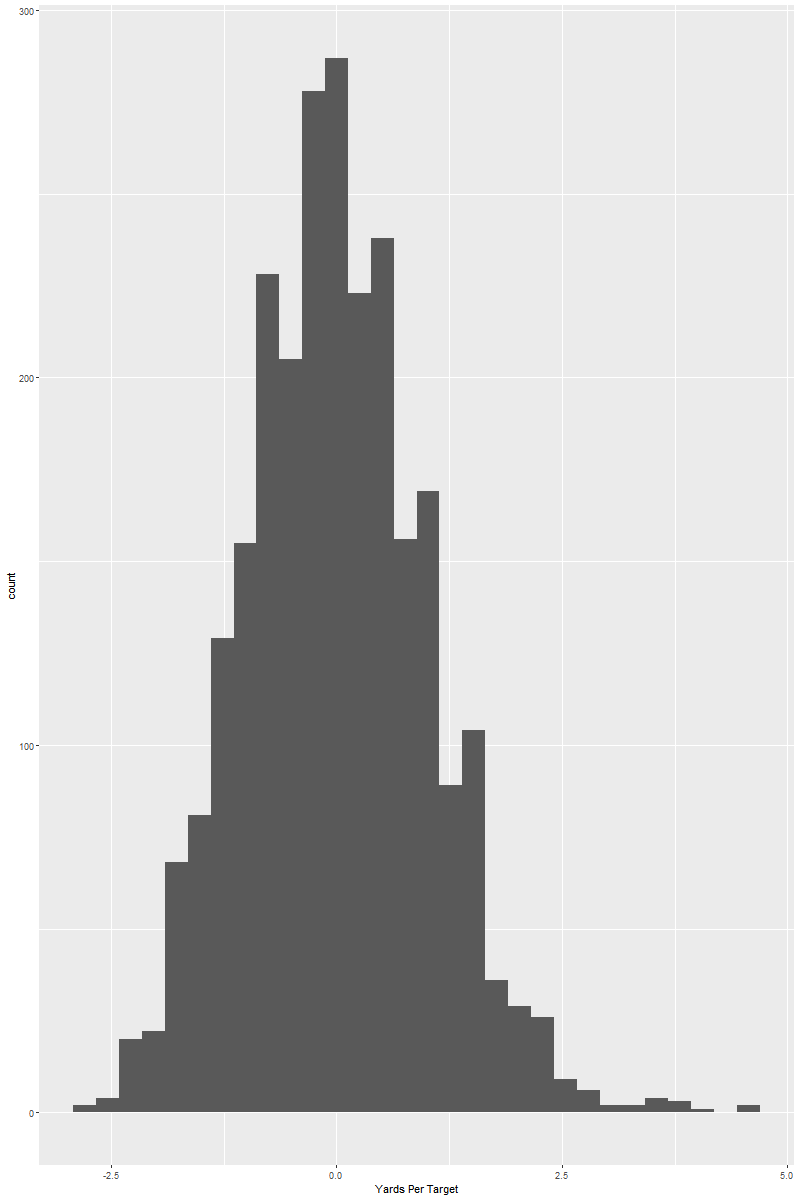
Appendix E: Average Yards Per Target Per Player (Univariate)

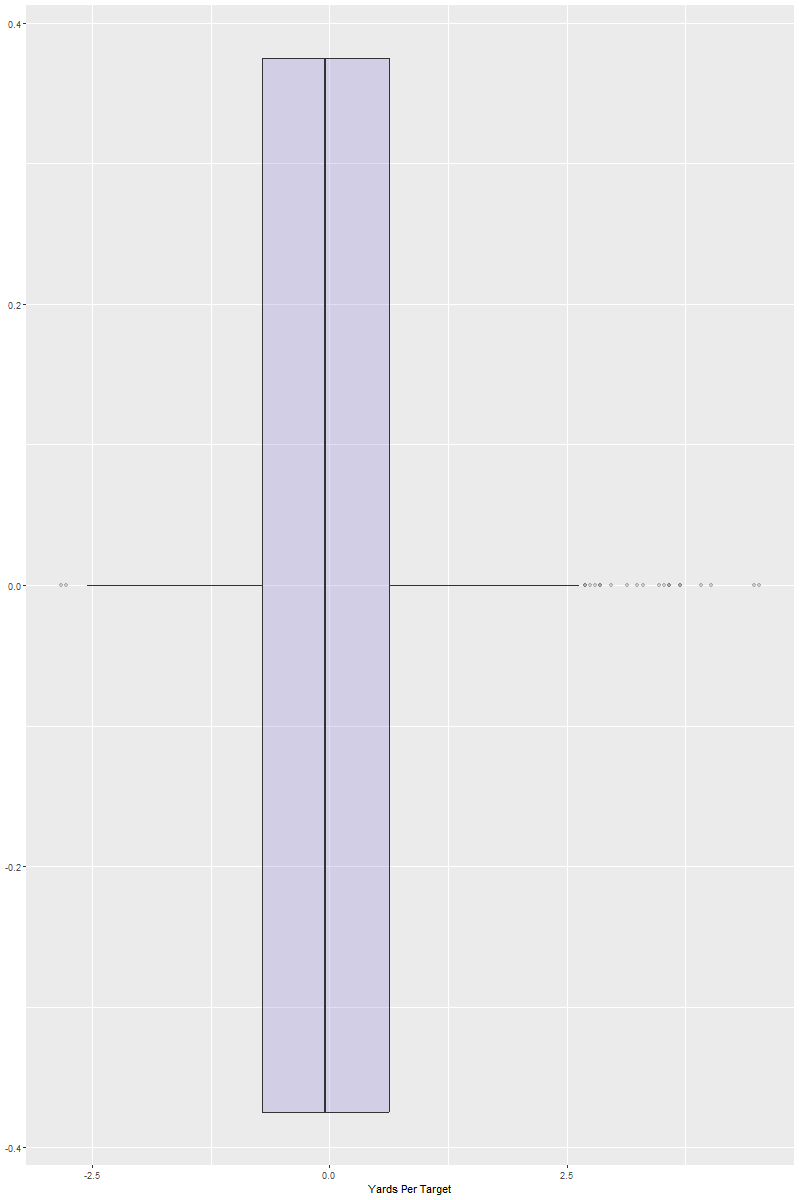
Raw Data:





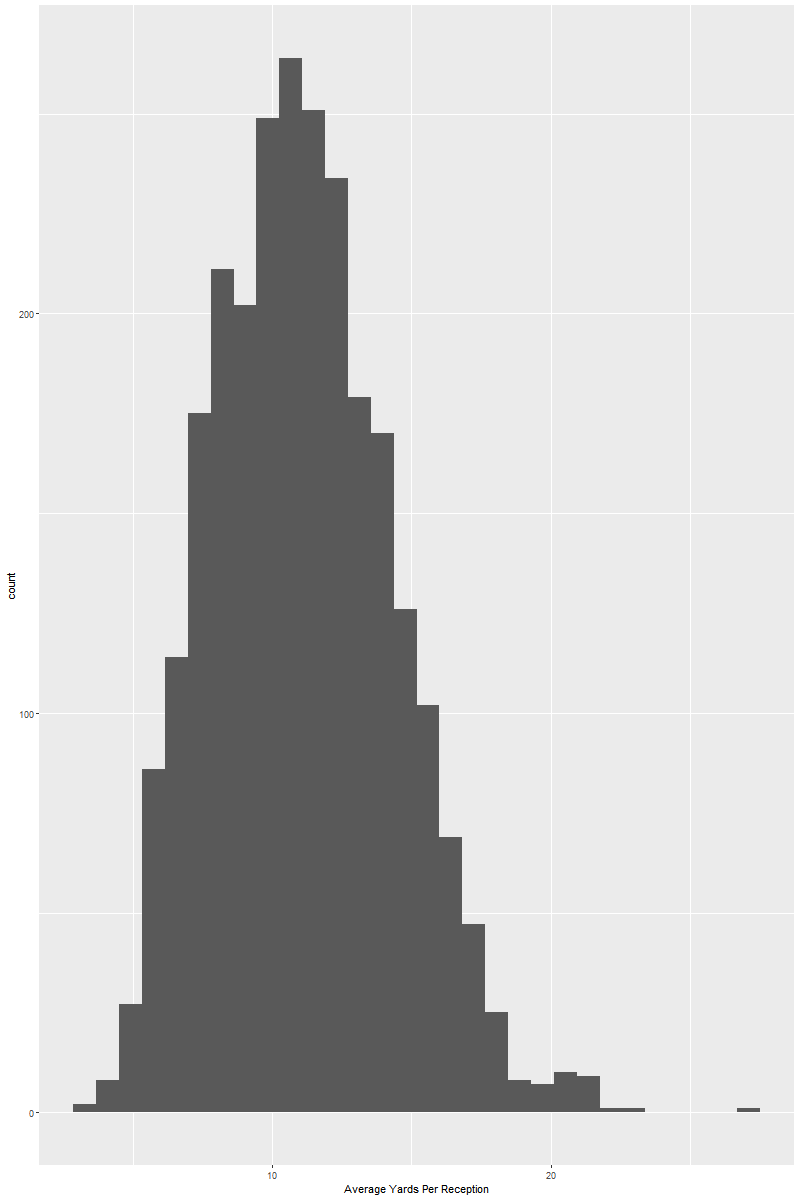
Standardized Data:

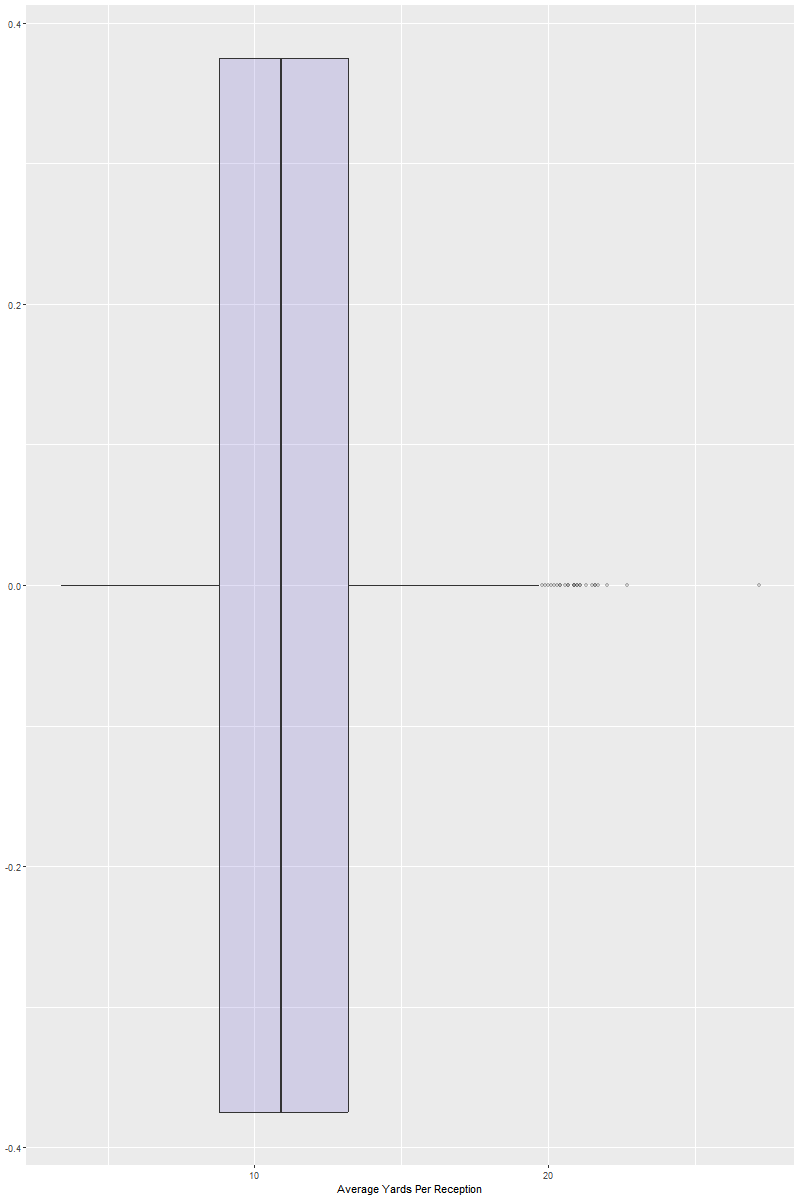




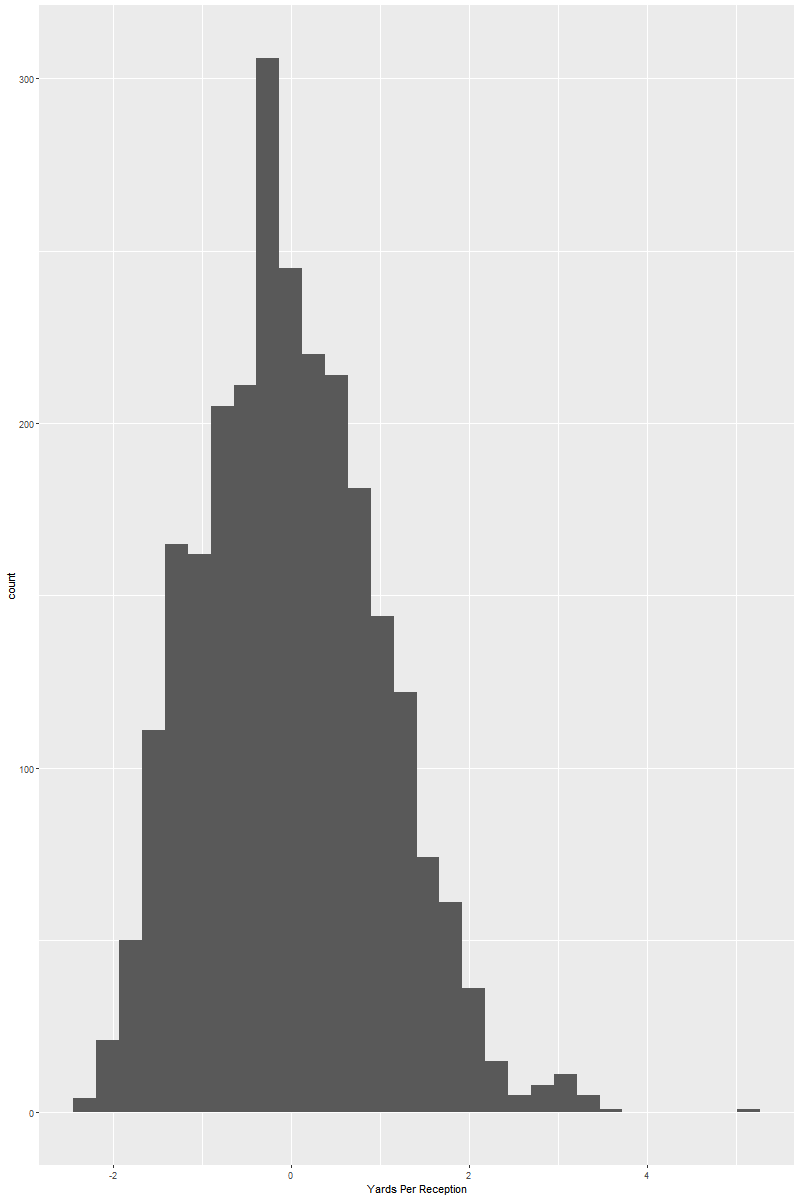
Appendix F: Yards Per Reception Per Player (Univariate)

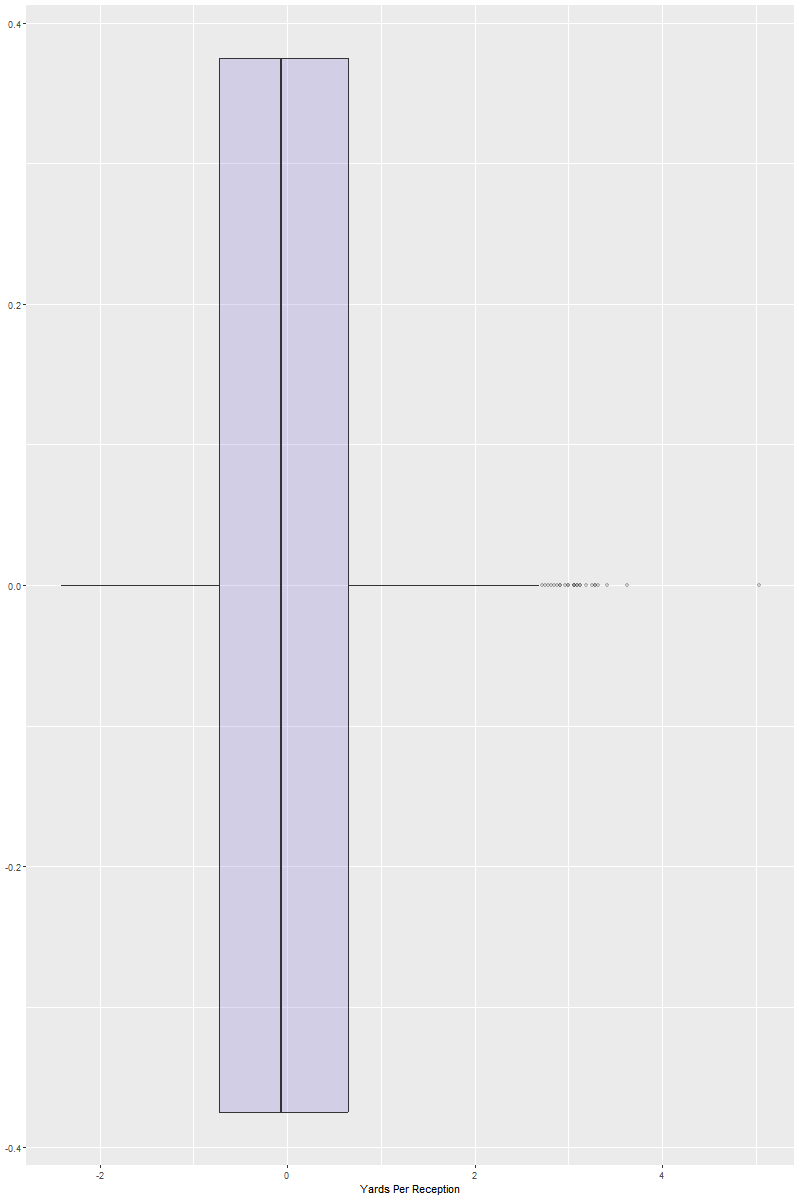
Raw Data:





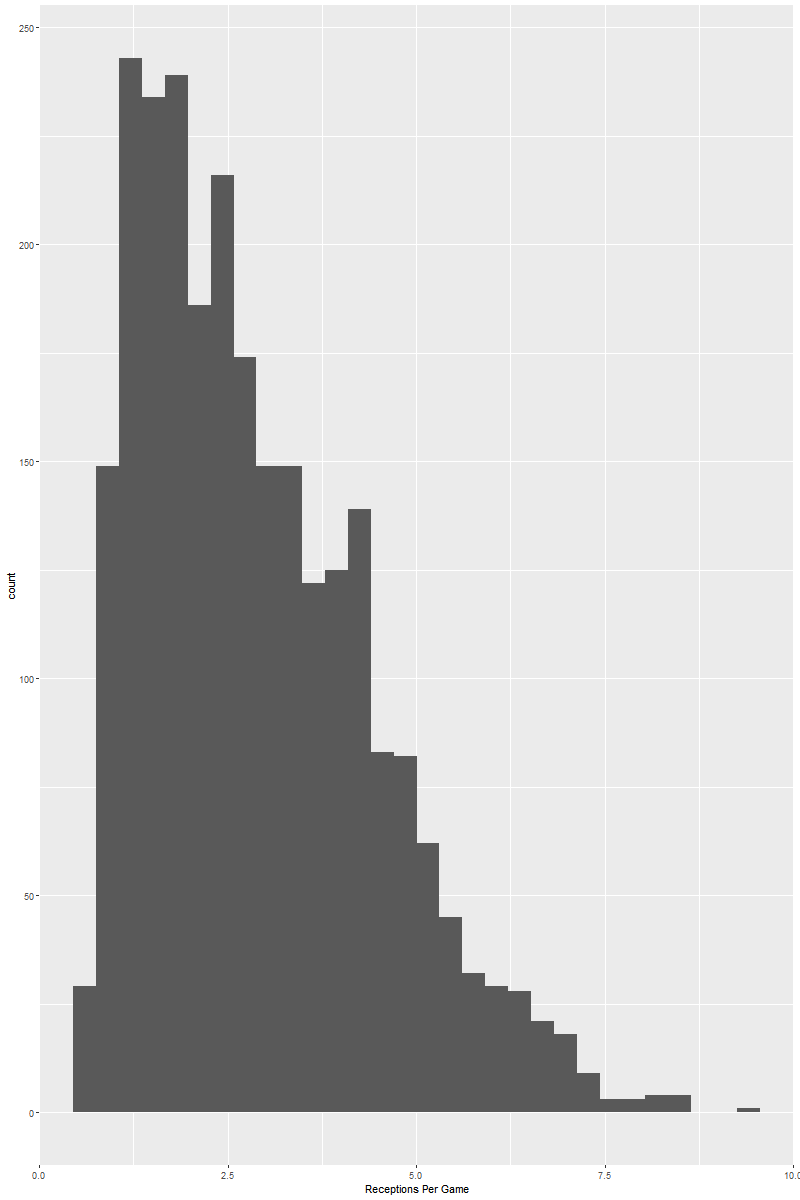
Standardized Data:

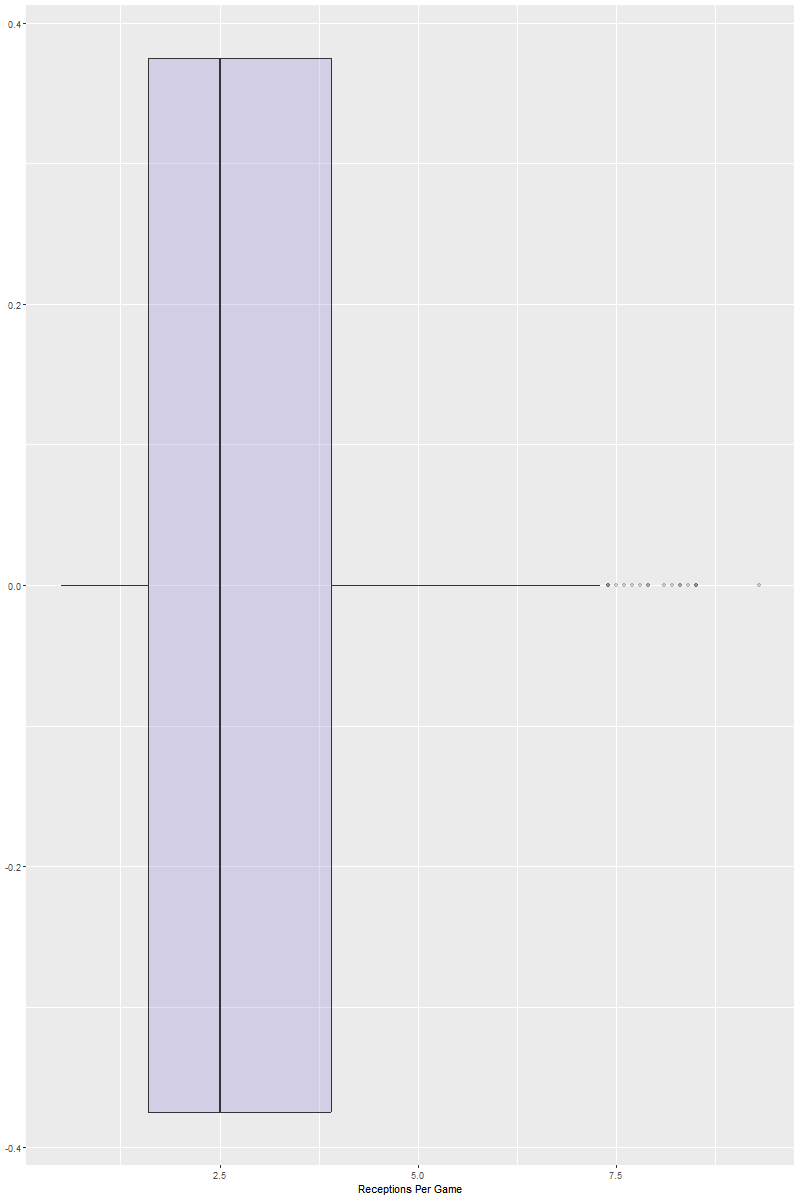




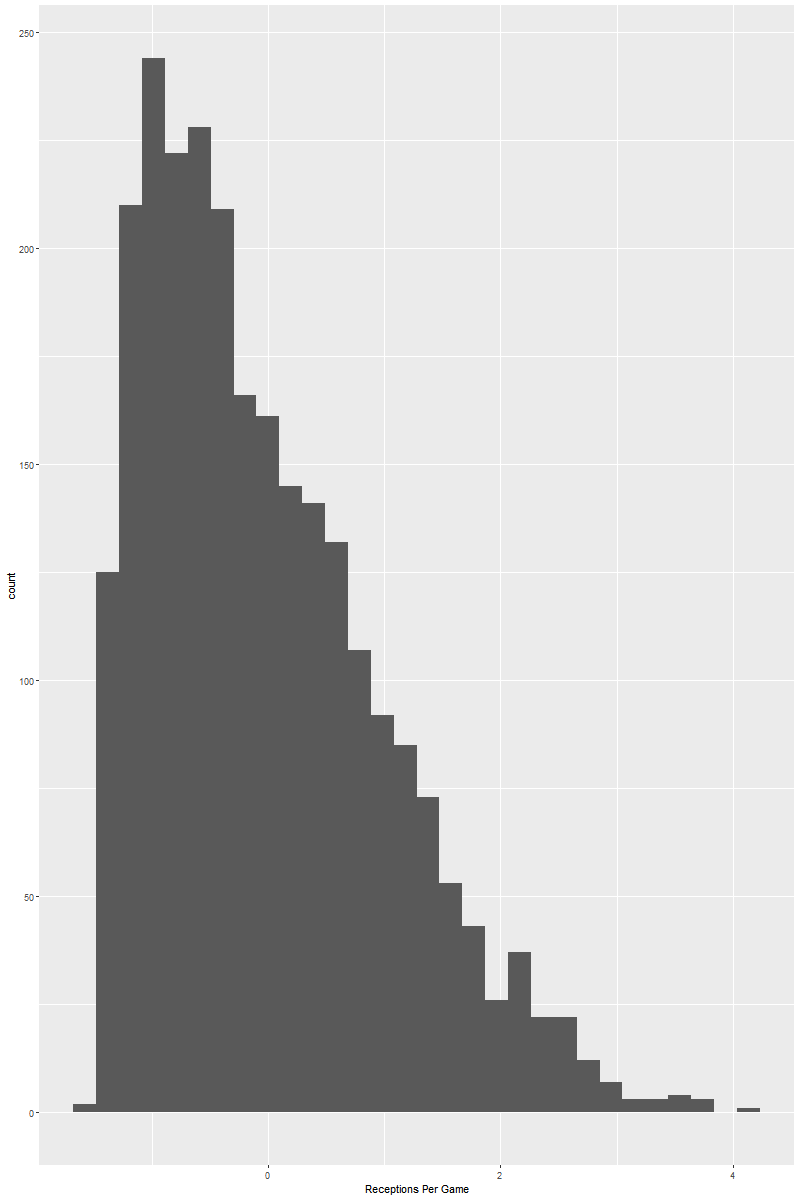
Appendix G: Average Number of Receptions a Game Per Player

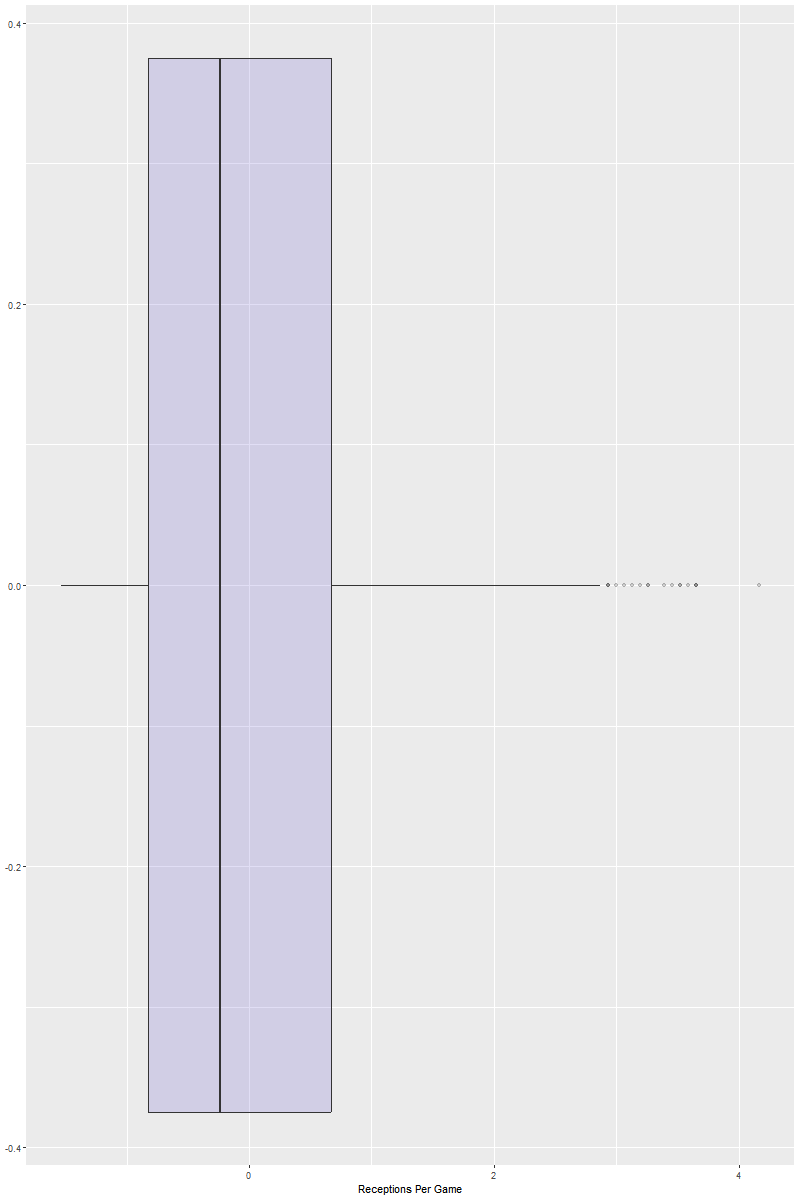
Raw Data:





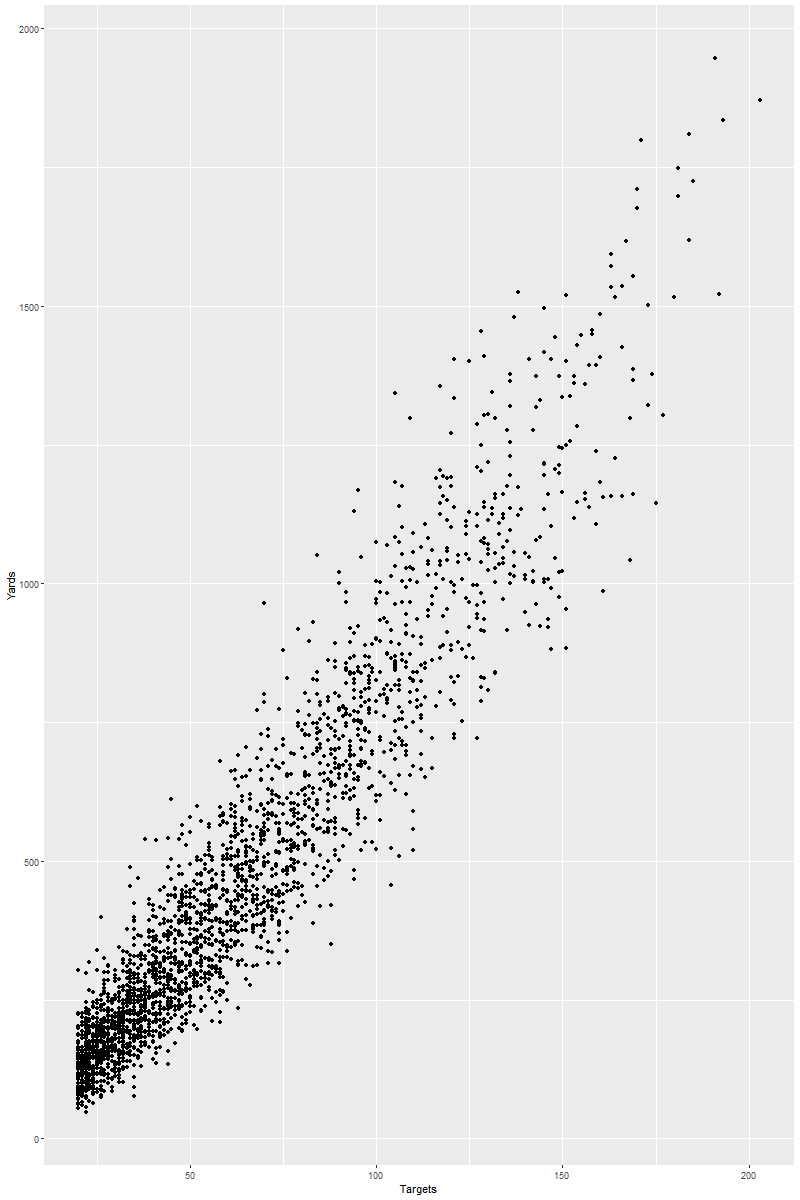
Standardized Data:



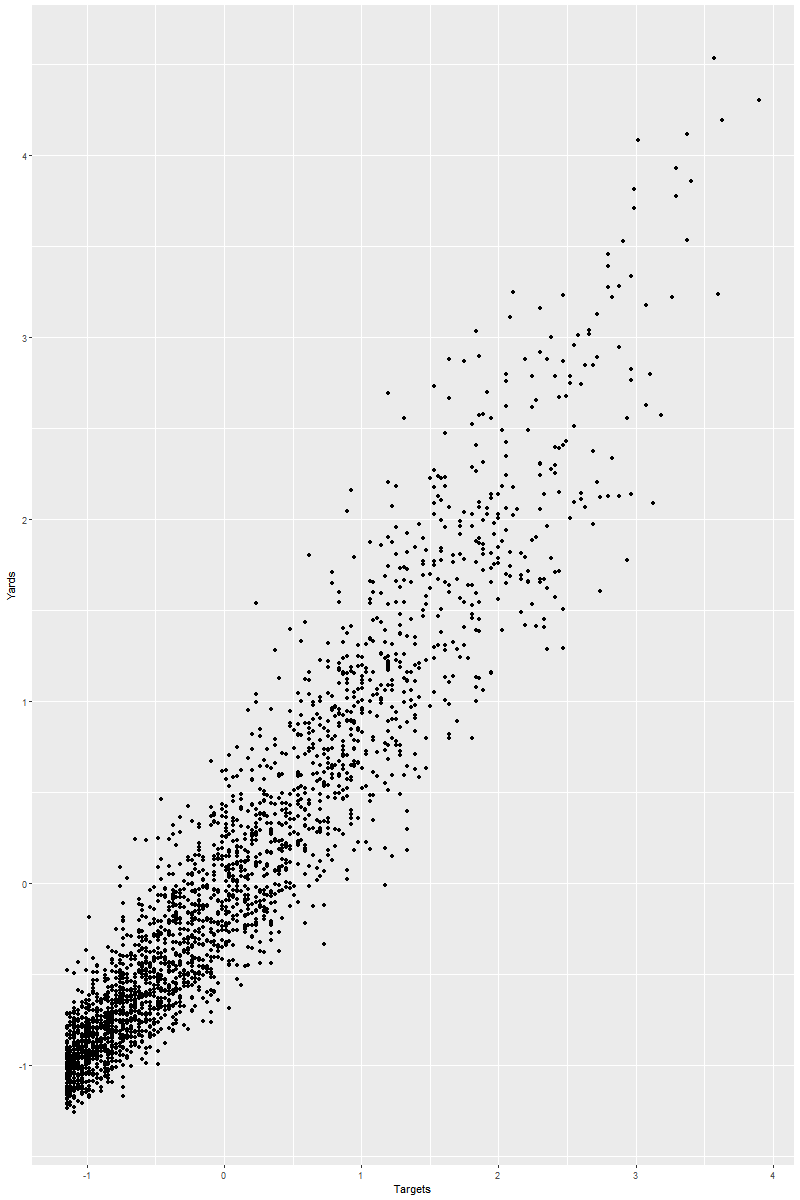


Appendix H: Targets vs. Yards (Bivariate)

Raw Data:

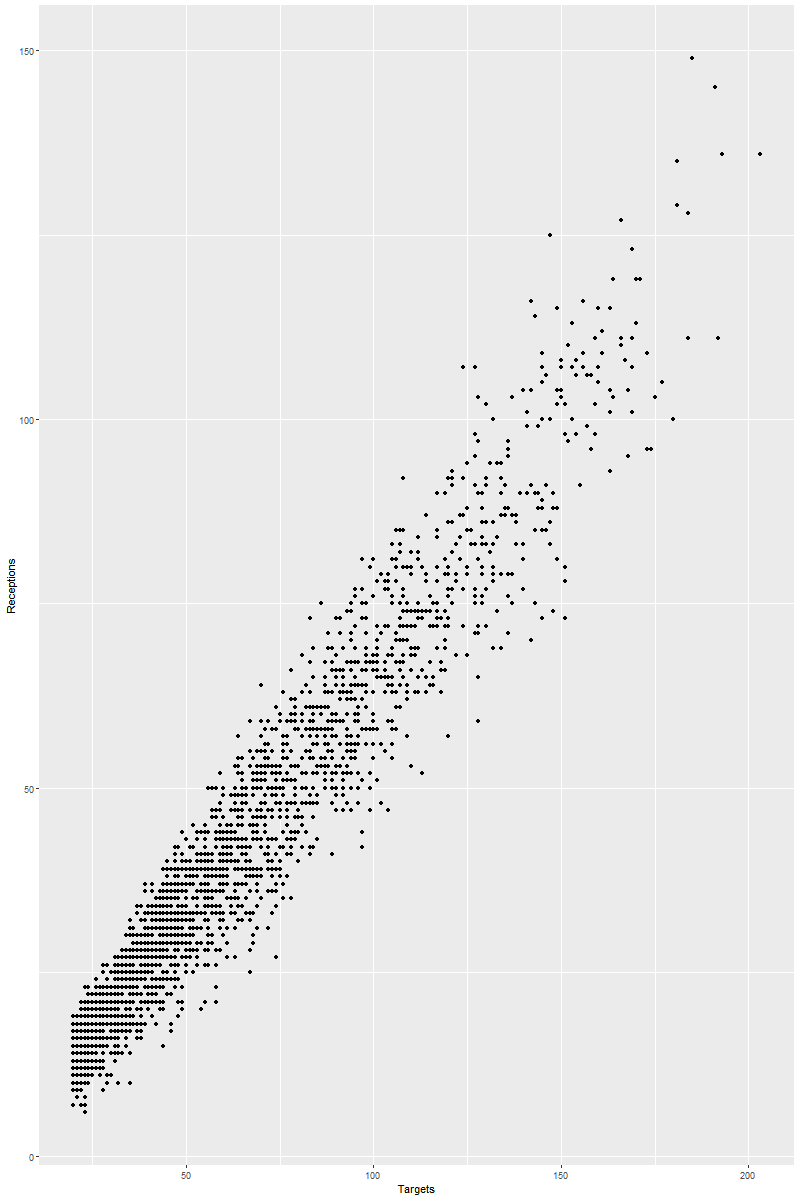


Standardized Data:

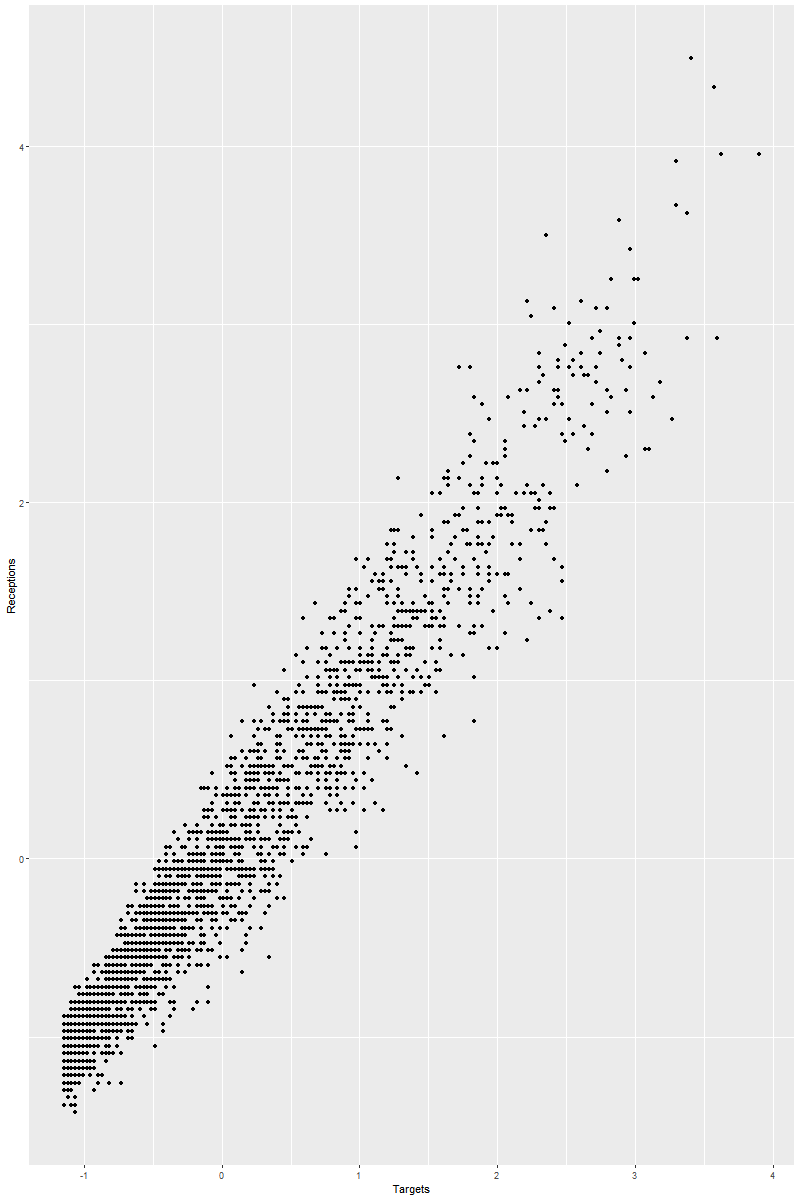


Appendix I: Targets vs. Receptions (Bivariate)

Raw Data:

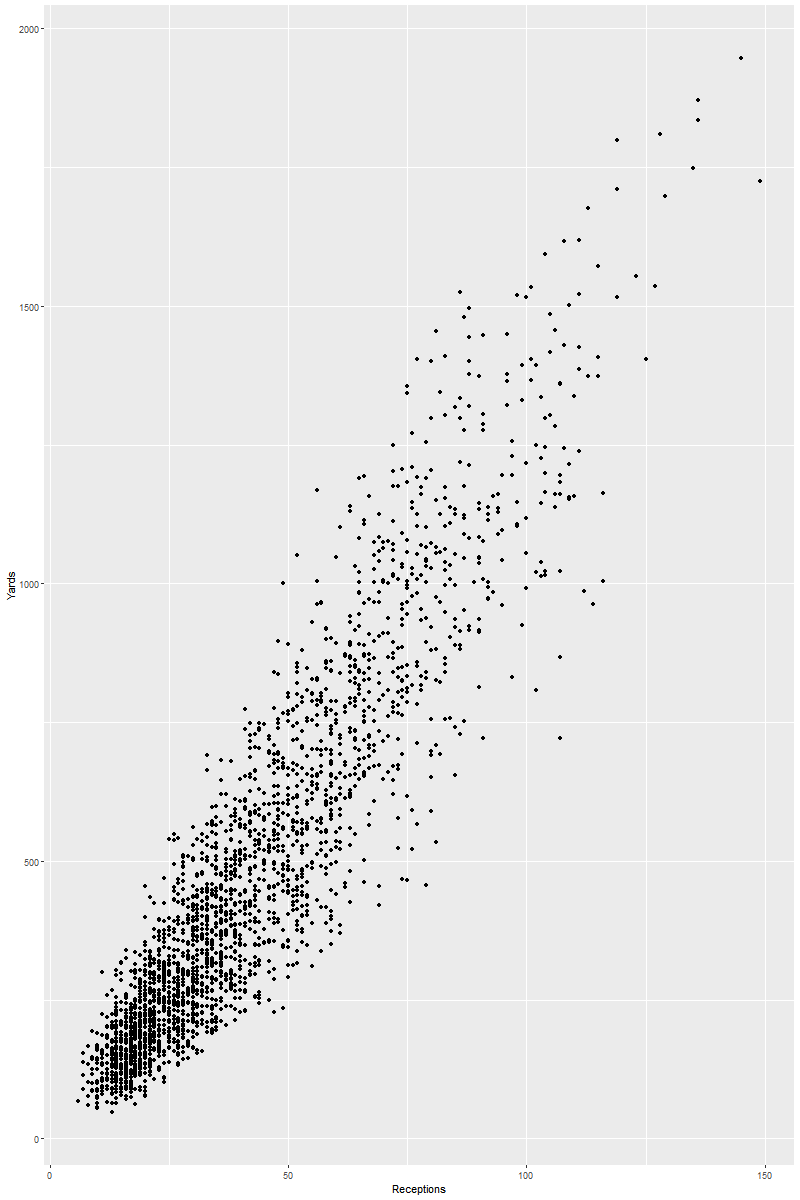


Standardized Data:

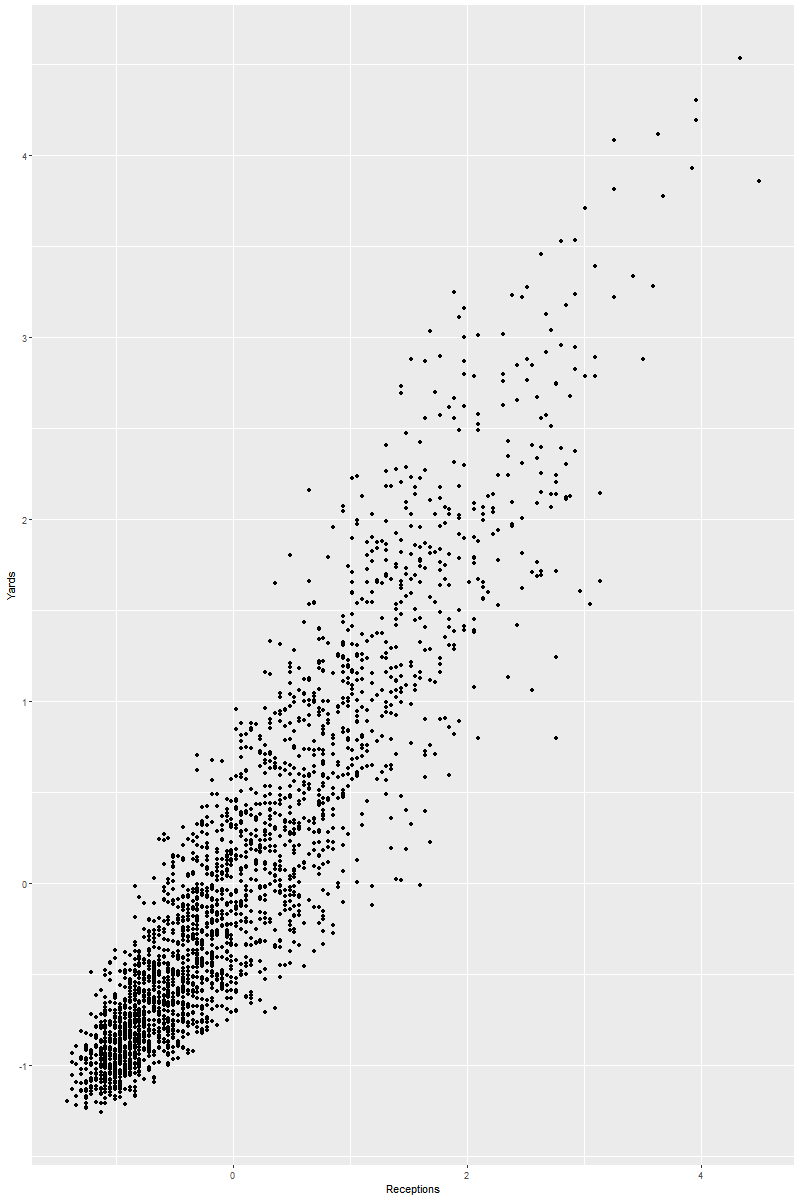


Appendix J: Receptions vs. Yards (Bivariate)

Raw Data:

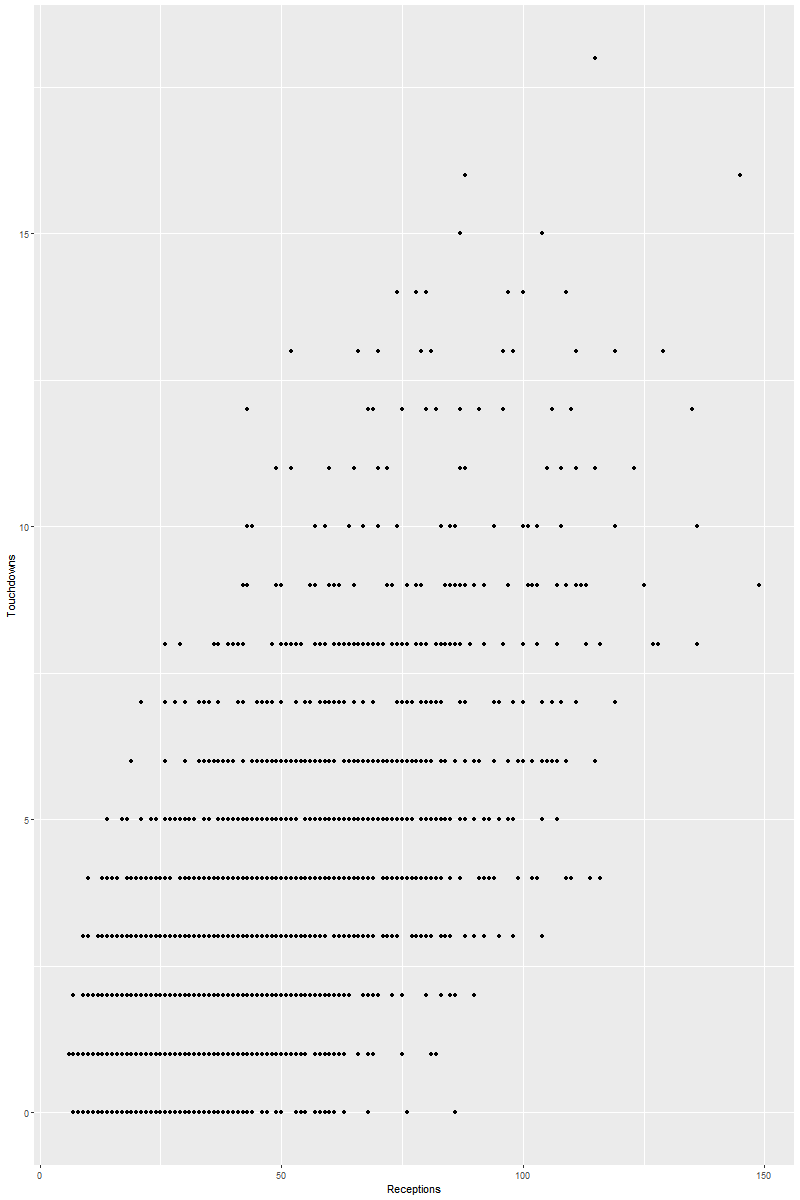


Standardized Data:

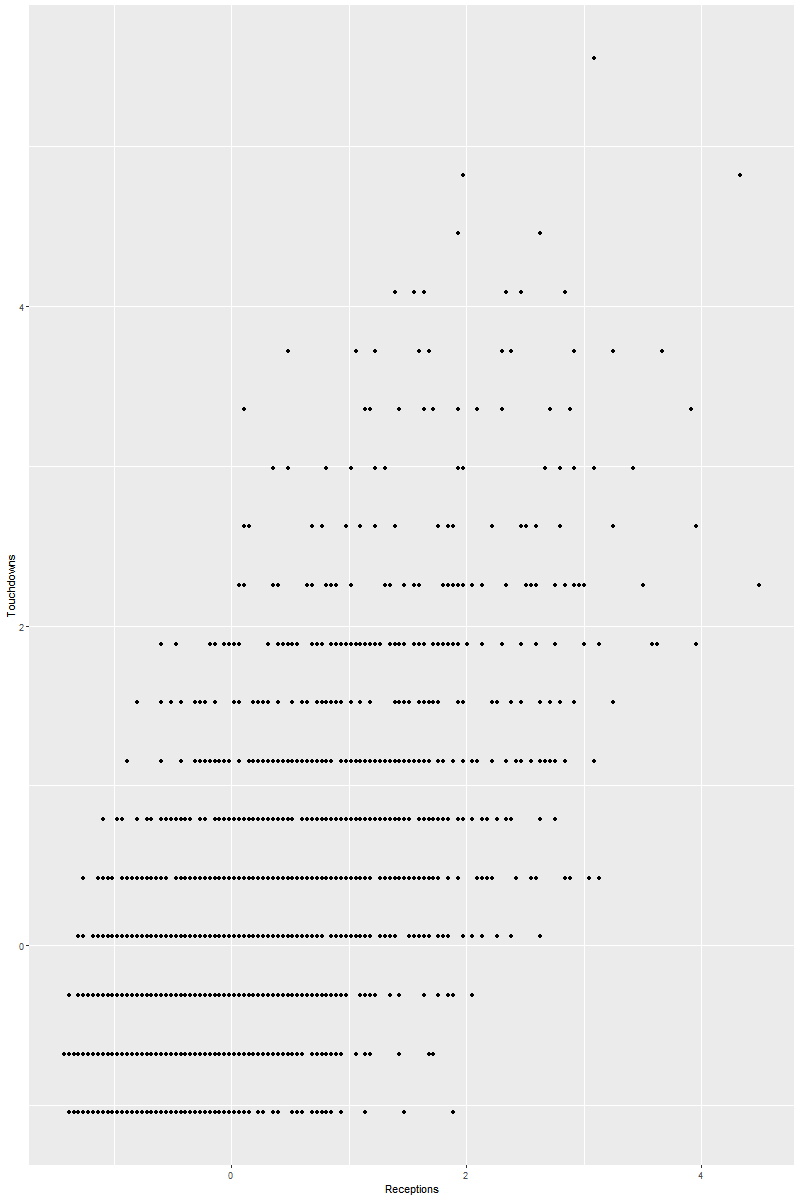


Appendix K: Receptions vs. Touchdowns (Bivariate)

Raw Data:

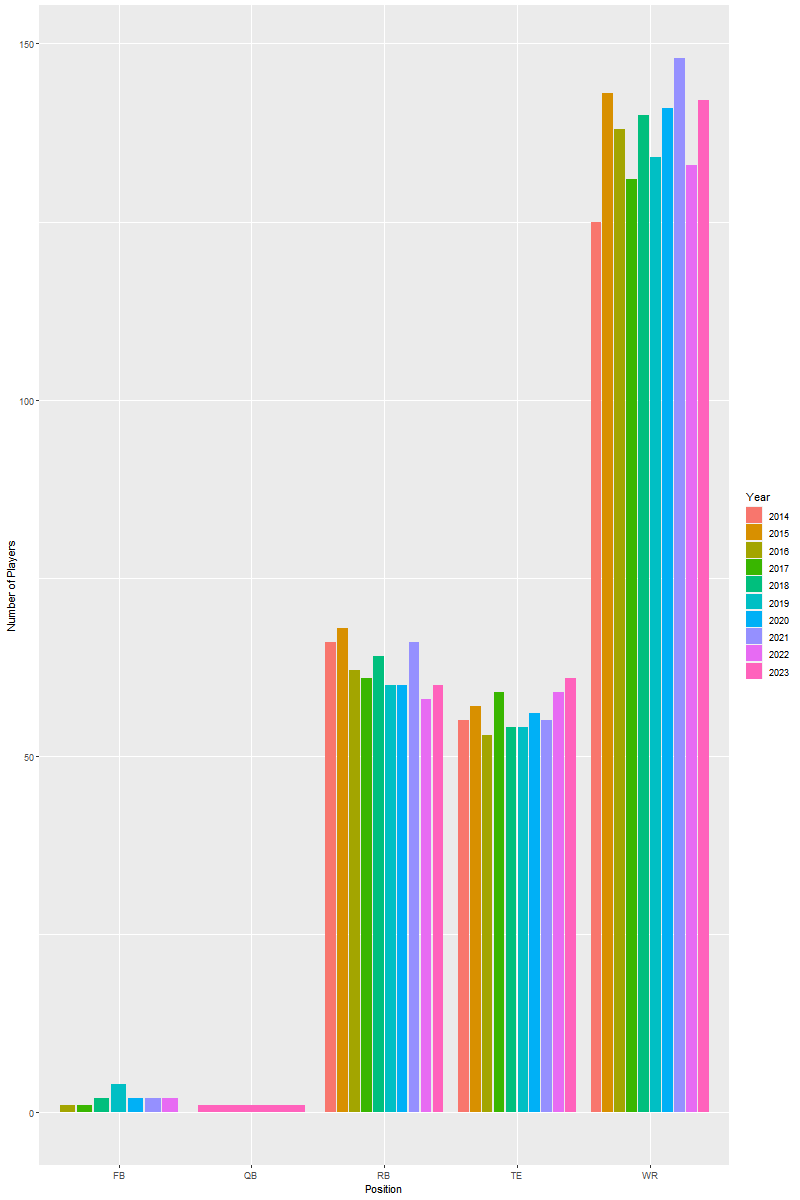


Standardized Data:

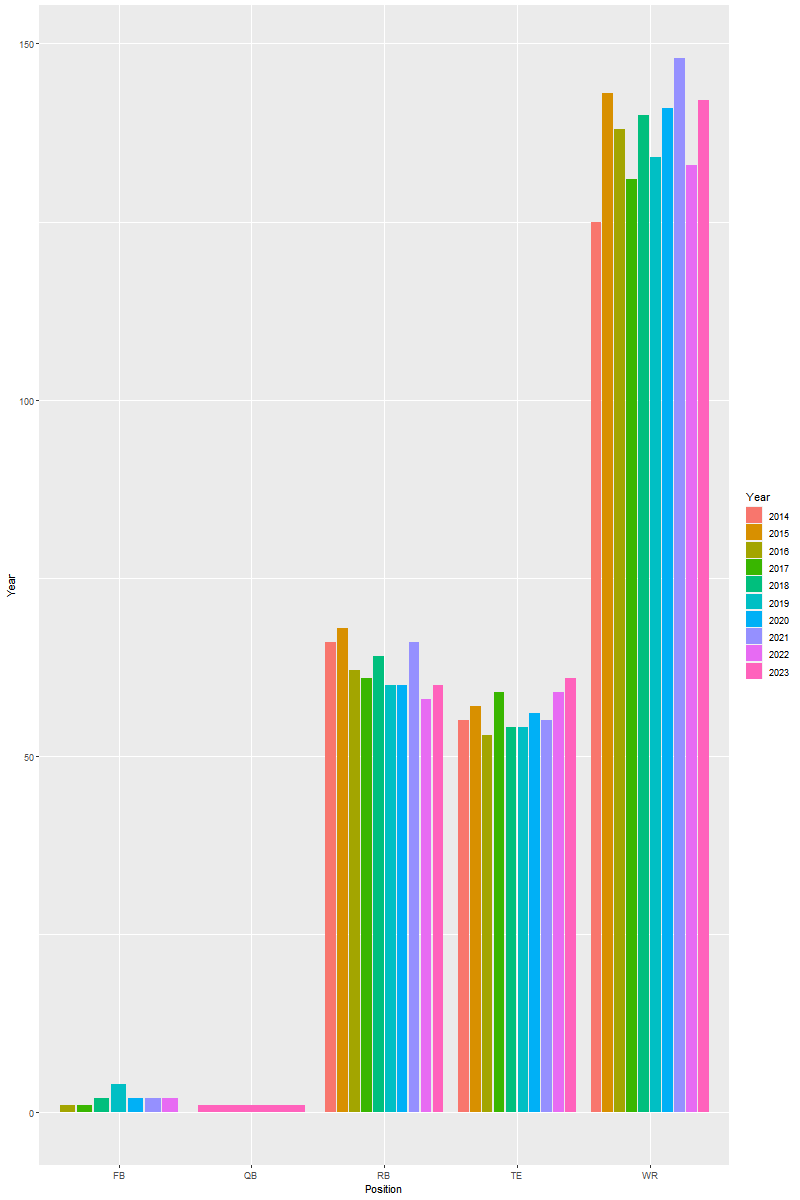


Appendix L: Number of Players Per Position Per Year (Bivariate)

Raw Data:

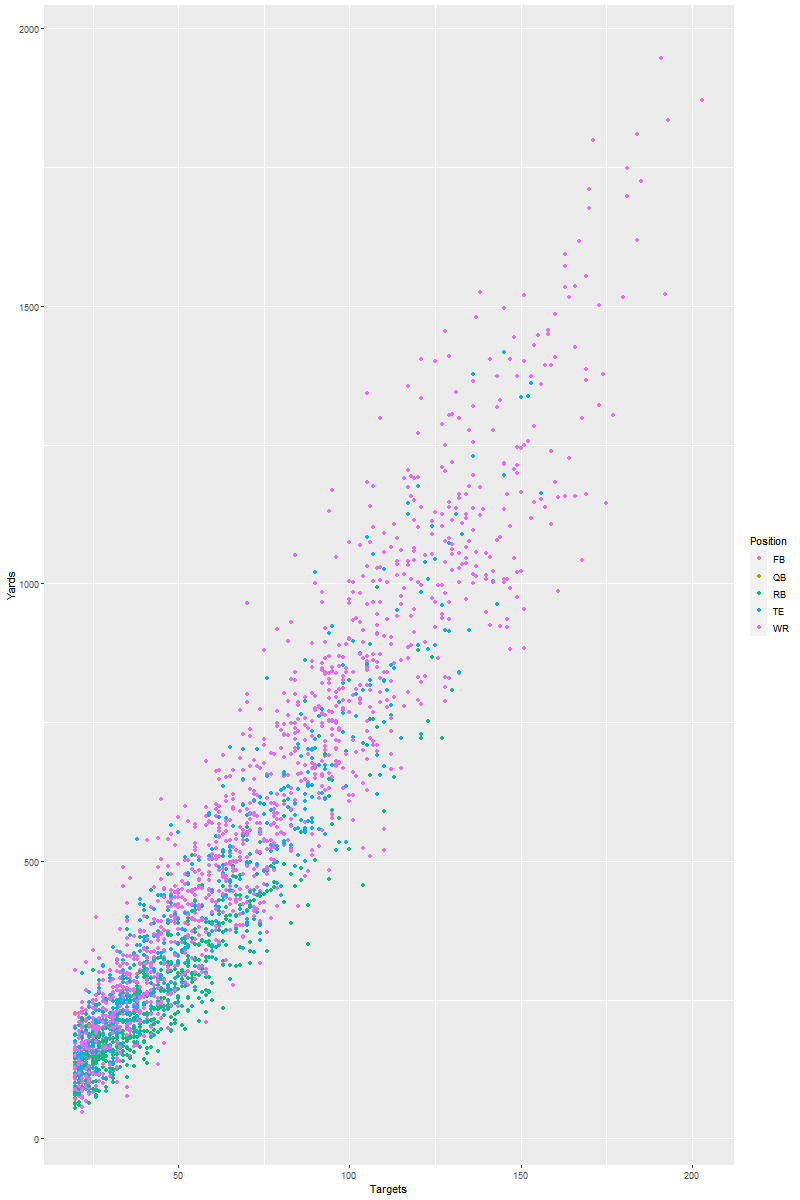


Standardized Data:

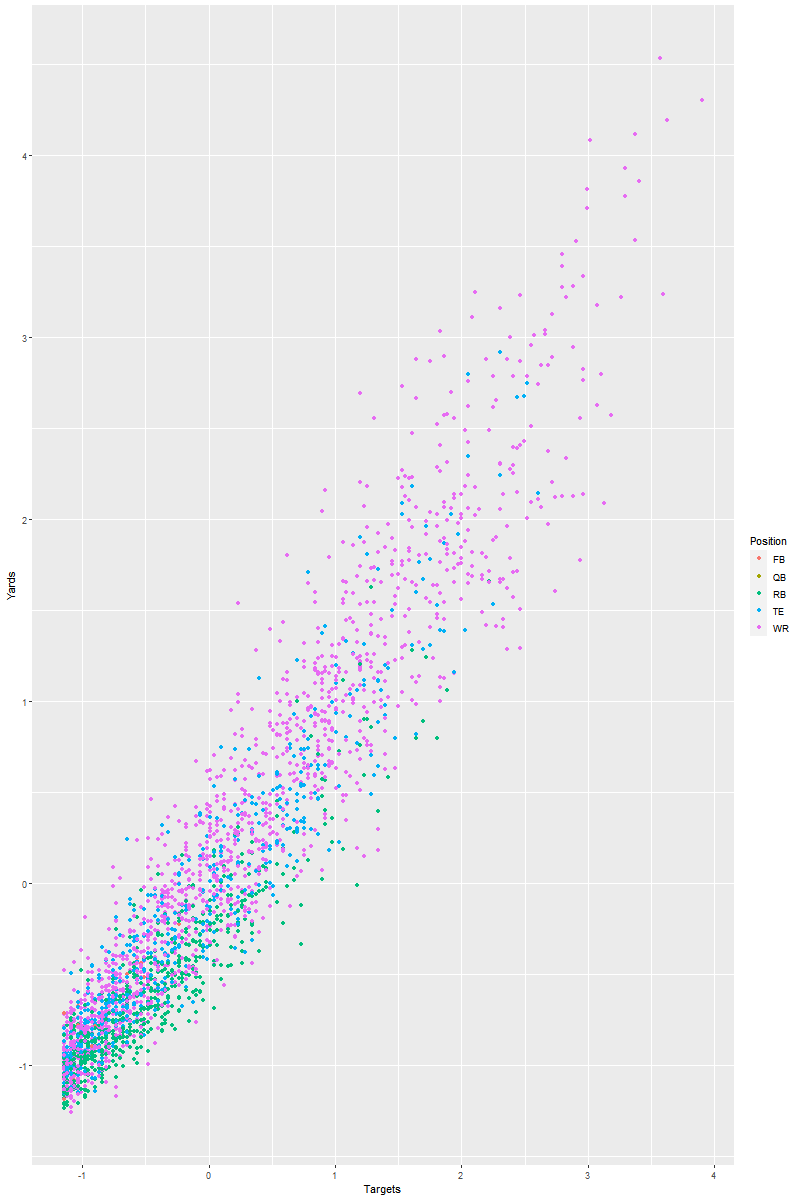


Appendix M: Targets Per Position vs. Yards (Multivariate)

Raw Data:



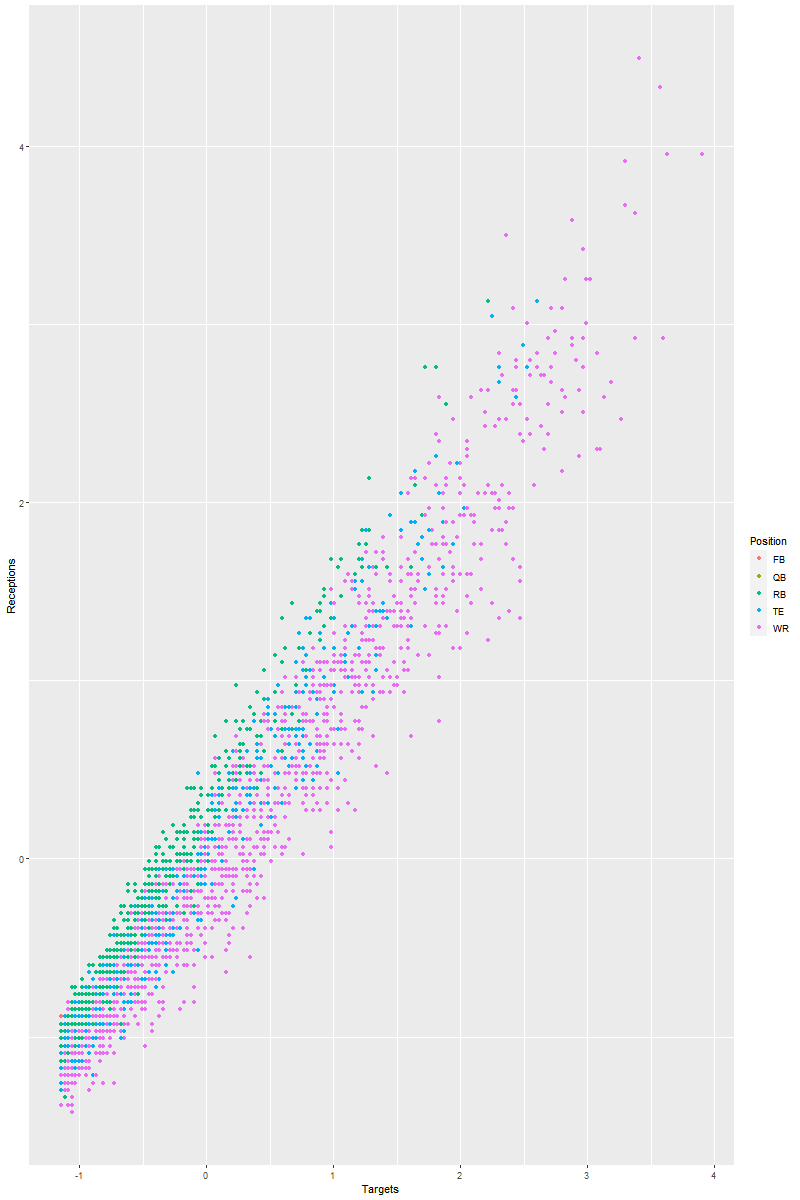
Standardized Data:



Appendix N: Targets Per Position vs. Receptions (Multivariate)

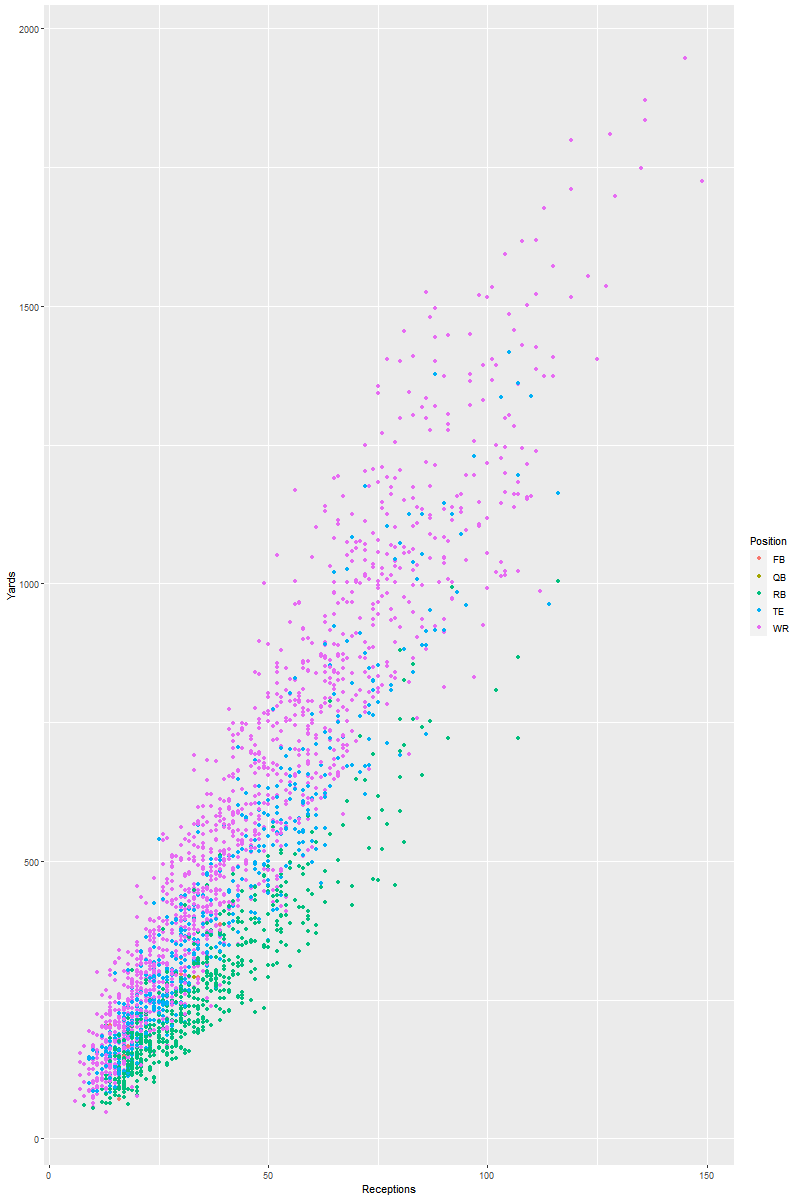
Raw Data:



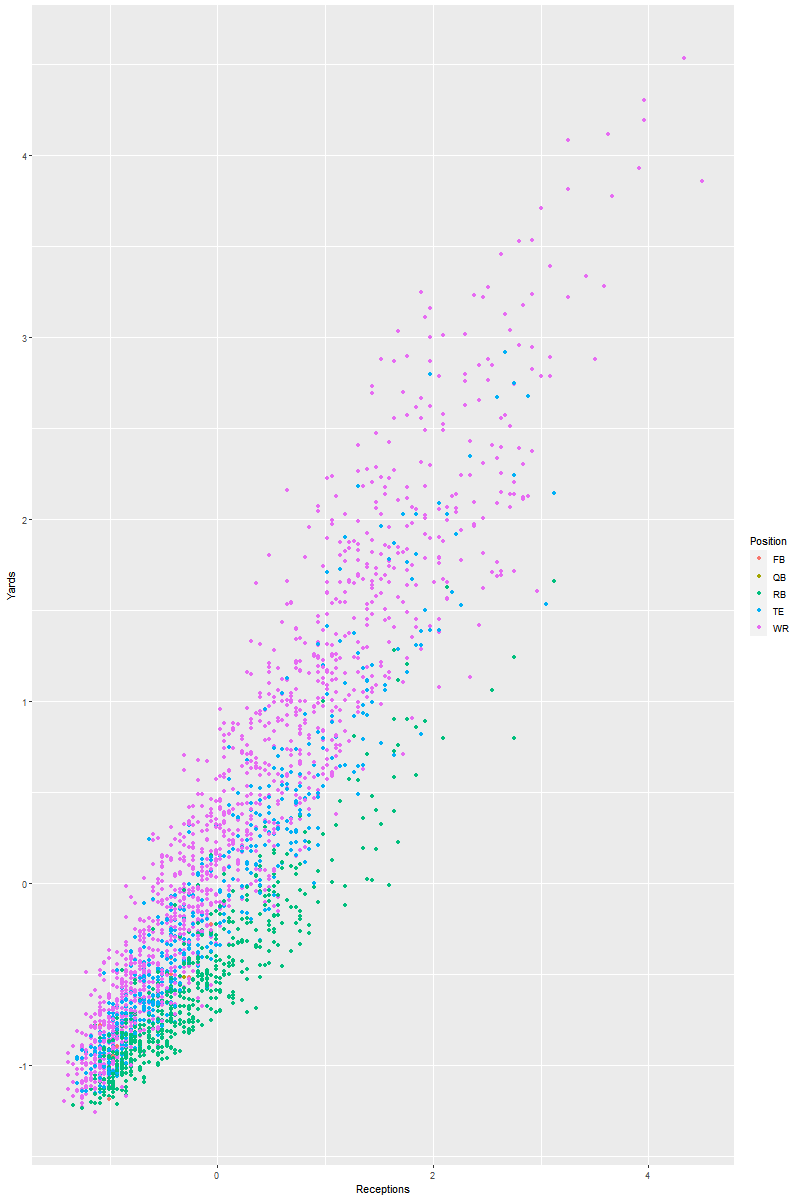
Standardized Data:

Appendix O: Receptions Per Position vs. Yards (Multivariate)

Raw Data:

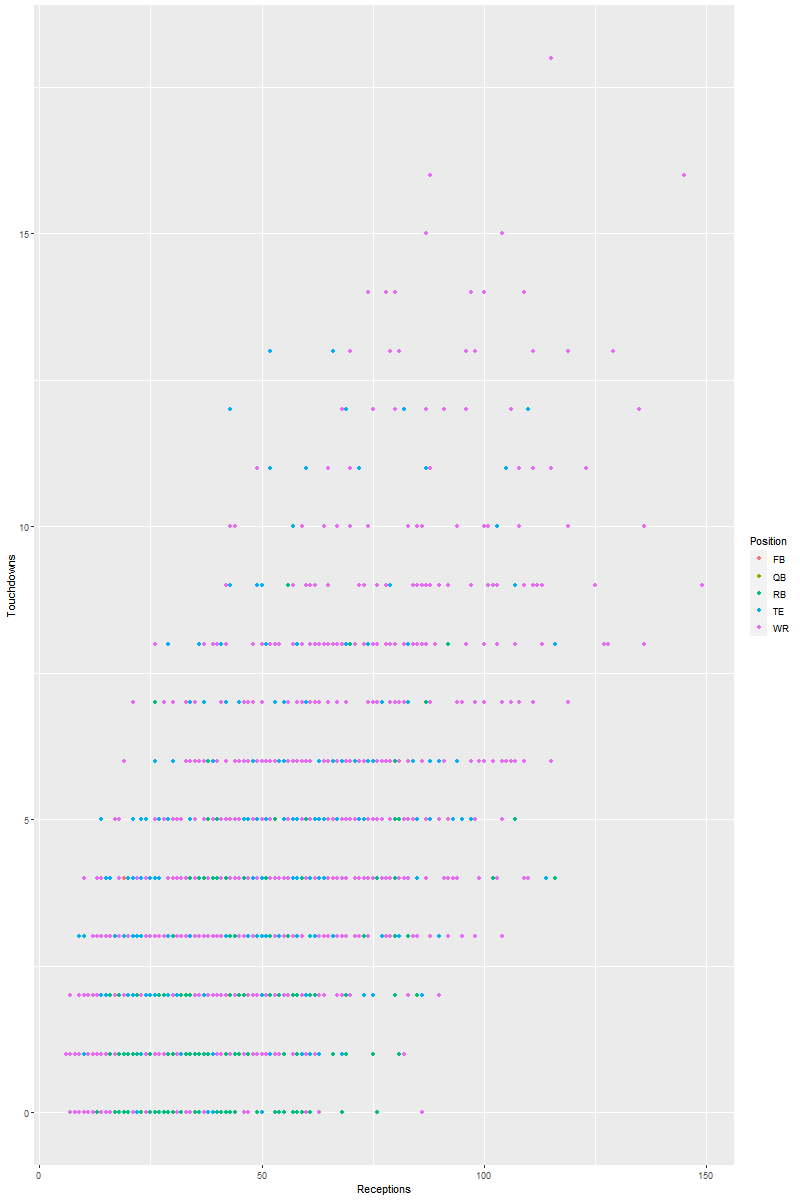


Standardized Data:



Appendix P: Receptions Per Position vs. Touchdowns (Multivariate)

Raw Data:



Standardized Data:

