Syracuse University

Hierarchical Clustering Analysis on Offensive Skill Positions in the NFL

By. Alejandro Pesantez

SAL 496

Rodney Paul

May 16th, 2021

Abstract:

This thesis looks to analyze NFL player season stats during the 2009-2017 NFL seasons. A hierarchical clustering analysis was done in order to cluster each offensive skill position (Quarterback, Running Back, and Wide Receiver), into different "cluster" groups, and to see which cluster for each positional group gives the best chance for a team to win. The data was collected from the nflscrapR GitHub which included passing, receiving, and rushing statistics from the 2009-2017 NFL seasons. Statistics that were used from each dataset were, passing yards, rushing yards, receiving yards, and other basic football statistics for the positions observed.

Table of Contents

Introduction:

Winning, that's the goal for every team in any sports league. Everyone benefits when their team wins, such as the fans, the front office, the players, and even the coaches. However, winning in football is a lot easier said than done. In order to win in the NFL, it takes a lot of pieces to come together. This is because football has twenty-two players on the field at all times. Since this is the case there are a lot of pieces that come together.

With there being that many players on the field at the same time, there are also a bunch of different player archetypes. On the offense there are quarterbacks that only pass the ball within the pocket. In other offenses a quarterback could be outside of the pocket more often and can also pick up yards on the ground. There are also running backs and wide receivers that have different archetypes as well. There are short yardage running backs that are used to get that extra one yard in important situations. Then there are versatile or dual threat running backs that can catch and run the ball effectively, which can really stretch defenses. Then for the receivers, there's an archetype that are red zone threats, where the receivers expose defenses when their tan is close to scoring. There's also a receiver archetype where receivers get open in the middle of the field or for short-to-mid yardage gains to help their team move down the field.

There are even archetypes for offensive lineman, linebackers, defensive backs, and so on. Since there are so many different archetypes, there are a lot of decisions GM's and coaches need to make in order to decide what archetype of the different positions gives their team the best chance to win. This thesis will go in depth about which archetypes are best for each offensive skill position. These positions include Quarterbacks, Running Backs, and Wide Receivers. Due to the lack of public advanced statistics on other positions in football, these positions were the most reasonable to do this analysis on.

In order to figure out the different archetypes for the offensive skill positions, a hierarchical clustering analysis was used in order to cluster the different positions into their perspective archetype groups. The quarterback, running back, and wide receiver data was from the nflscrapR package [30], and once this data was clustered, different player archetypes were made for each offensive skill position. Once the archetypes were made, each archetype was compared to the average actual winning percentage and the average WPA (win probability added) of the archetype in order to see which archetype is best for helping a team win.

The analysis is broken up into five sections. It begins with a literature review of some papers and articles that are referenced and that helped come up with the idea of this thesis. The next section is a summary of the data used. This section discusses where the data is from, and what variables the data has. The ensuing section is the methodology used for the analysis, where it goes in depth of the methods used on the data in order to come up with the optimal number of clusters needed for each of the positional datasets. The following section goes in depth about what the clusters results show. This section gives visualizations of summaries of the statistics for the different archetypes created by the clusters. This section also shows the different clusters compared to the WPA and actual winning percentage to show which archetype is best for winning at each position. The last section contains the conclusions and final results of this analysis, as well as what improvements could be done to this thesis to get more meaningful determinations.

Literature Review:

A. Clustering Techniques

In this article, Blake Atkinson talks about how he classified and made different NFL player styles. He was always very intrigued in the Tiki Barber and Ron Dayne dynamic and really wanted to see if he could show this using statistics and clusters. In this article he uses principal component analysis along with clustering algorithms in order to easily sort out players into separate categories. Throughout the article he gets different clusters for the three different position groups he uses. For each of the position groups he got six different groups to compare from, while using 11 variables to base them off of. In the end he wanted to compare Barber and Dayne one more time and looked their rushing vs. receiving efficiency. He found that Ron Dayne was efficient at catching the ball, which was surprising, however, as time went on, Ron Dayne trended toward league average and Tiki Barber turned into a monster. [4]

In the article DaSilva does an analysis of the importance of each NFL position. It shows the average salaries of these positions and then ranks them. Some surprising things that I found in the article is that he put running backs 10 out of 15 which was very interesting to me. I figured they'd be much higher but when you think about it, they are pretty replaceable if you have a good offensive line. Another ranking that shocked me was the number 2 position they had. At the number 2 position are pass rushers, which I thought would be high on the list but definitely not #2. This could be very useful for my project because when generating a "perfect lineup" I am going to have to weight positions differently, so this article could be very useful to base these position weights off of. [13]

In this article it goes through a very well put together lineup analysis just like one I want to do but in the NBA. An interesting thing that I saw when they were coming up with variables they were going to use for clustering, was that they used height. This had me thinking that physical statistics would definitely be important for my analysis as well, except it would be different for each position. They first used a k-means to cluster but didn't like their results, so then they decided to use a gaussian mixture model instead. They ended up getting 9 different clusters and the positional names were spot on given what type of player they were. Once they got the different type of player, they then had to decide what stat to use in order to compare lineups, in this case they used net rating. Then with that net rating they converted it to a Bayesian net rating to counteract noisiness of the lineup net rating. They then constructed soft lineups based off nine predictor variables. They then went into deeper analysis using regressions, and random forest models in order to compare and build these different line ups. I feel like this article could really help me start creating a plan on how to come up with the proper methodology for me to do this lineup analysis in football. [8]

In this article the author Kassambara talks about different ways to cluster in R. This will be helpful for my thesis because I'd probably like to try different ways of clustering to see what result I like best or makes the most sense. The beginning of the article talks about what packages to load when beginning to cluster which are cluster, factoextra, and margitir. It then goes through on how to prepare your data so you can cleanly cluster. The first method the article goes over is partition clustering which subdivides the data into a set of k groups. The next way the articles suggests is hierarchical clustering which is very similar to partitioning except for the fact it does

not require you to pre-specify the number of clusters to be generated. It then even goes into clustering validation and evaluation. This will be super useful because it shows how to determine the optimal number of clusters and shows you how to cluster validation statistics. Overall, this article is going to be crucial for my thesis because it will give me the techniques and setup of clustering with my data in R. [2]

In this article George talks about what he thinks are the 5 best clustering algorithms that data scientist needs to know. The first way of clustering he starts talking about is K-means clustering which he says is the most well-known clustering algorithm. He goes in depth on how to set up the algorithm and how to interpret the graphs. The next clustering method he talks about is mean-shift clustering. He says that mean shift clustering is a sliding-window-based algorithm that attempts to find dense areas of data points. He said that it is a centroid-based algorithm meaning that the goal is to locate the center points of each group/class, which works by updating candidates for center points to be the mean of the points within the sliding-window. He then when on to talk about Density-Based Spatial Clustering of Applications with noise (DBSCAN). He says that DBSCAN is a density-based clustered algorithm similar to mean-shift, but with a couple of notable advantages. The advantages are that it does not require a pe-set number of clusters at all. It also identifies outliers as noises, unlike mean-shift which simply throws them into a cluster even if the data point is very different. Additionally, it can find arbitrarily sized and arbitrarily shaped clusters quite well. Then the last two algorithm he talks about are hierarchical clustering and expectation maximization using Gaussian mixture models which I've summarized in other articles above. He says that using GM Is more flexible than k-means and that hierarchical is good because you don't have to specify the number of clusters at the beginning you can even select the number of clusters that look best in the tree. Overall, this article is going to be useful because it shows me very unique and different ways to cluster with my data. [31]

In this article Sam talks about clustering using soccer analytics. He talks about what exactly he wanted to cluster. In this case he just wanted to cluster forwards in soccer. Since this was the case, he removed stats from his data like Successful Header Percentage in The Defensive Third, Tackles Made in the Defensive Third, as well as a few others that are defensive in nature and are not pertinent in measuring a forward's overall ability. He then shows the stats that he did end up doing which were forward based such as pass percentage, shots on target, goals, etc. After selecting the statistics, he then needed to figure out how many groups to cluster. In this case he used a thing called the elbow method to figure out how many groups to cluster. Then he plugged in the metrics into an algorithm that performs k-means clustering. He noticed in the first cluster which was considered the "best" forwards, some very notable names were left out. He then adjusted his stats to make the cluster have better finishers. He concludes the article by saying clustering, an unsupervised form of machine learning, is an excellent way to help a team find the next diamond in the rough for a team in the player recruitment process. [17]

In this article Lee clusters Fifa 20 players to see which players are similar. In the beginning he briefly describes what clustering is and how it works. He then goes into specifics about k-means clustering. Which he pretty much says is specifying the k-means cluster, iterate until the cluster assignments stop changing, calculate the cluster mean for each k cluster, and then to proceed through the list of observations and assign an observation to the cluster whose

mean is nearest. He then goes through the data he is using which is basically FIFA ratings as well as the physical features of the players as well. He then filters the data so that he only has players that are 86 overall or higher. He then goes into what stats he uses for what and the graph produces 5 pretty nice clusters. You can see in the graph that even though some players in certain clusters are different positions, they are still very similar in play style, which could be very useful to use in order to create a an "ultimate team". Not only would this be useful for video games such as FIFA to try to make the best team, but it could also be used for real life soccer teams as well. [24]

In this article Haider talks about how he uses k-means clustering to define NBA positions and roster construction. He begins the article by recapping his dataset, which was every statistic from basketball reference during the years 2011 to 2018. To begin his clustering, he wants to do some exploratory data analysis and made some cool visuals with them as well. He then showed the data science methods he used which was principal component analysis (PCA), K-means clustering, using elbow method and silhouette scores, and scaling features. In this case he got 9 different clusters of types of players ranging from perimeter scorer, superstar, all-star, 3&D, etc. He then made so very interesting visuals of the clusters he made, from three dimensional graphs to two dimensional graphs. He then went in depth about what each type of player was, which I thought was really cool. For each type he took the average of the basic box score stats so you could look at the differences between each cluster. I definitely will want to look back on how he organized this so I could use it for mine. To end the article, he tried to answer two questions which were, "What's the difference between elite teams and average teams when it comes to player roles/styles?" and "Do winning teams have more or less players with a specific role/style? Does roster diversity play a part in winning?". He answered these questions by looking at his clusters. He said that overall, NBA finalist teams have more star power, and their inside players have a reserved role. Which he said that in turn, average NBA teams have less star power and rely on star inside bigs as their focal points. He finishes by saying that teams that want to win should have their roster to be more reflective of the pie chart for NBA finalist teams. [19]

In this article Anders talks about the basics of clustering but with using gaming analytics which I thought would be interesting to look at. In the beginning, he starts talking about what a behavioral dataset consists of. Which he says that they can be very big, time-dependent, and high-dimensional. He then goes on to say clustering is imminently useful for categorizing your players and for getting an overall idea about the variance in player behavior and how behavior is organized, but also for detailed analysis. He then says to present everything cluster analysis can do is vastly out of scope for his post. He then says reducing dimensionality and finding patterns are huge for clustering. This I think will be very helpful for me to use when trying to figure out what to cluster with my data. Lastly, he talks about the foundation of cluster analysis and how clusters work. Overall, I think this article will be useful in order to define what a cluster is. It will also be helpful to follow the step-by-step process he uses to set up his clusters. [3]

In this article Justin talks about how he decided to cluster NBA positions. In this case he used to use a statistical clustering tool to group players together based on the shots they've guarded according to data from stats.NBA.com. To begin his methodology, he decided to pull a list of 2015 players who were listed at the nearest defender for at least one shot and then he calculated the average height and weight for the shooters that particular defender defended. Then

he used k-means clustering in R. He said that in order to get more granular results he just computed the distance from the centroid. He said that this worked pretty well and that he regressed the stats to the mean so a player with only one field goal defended would mostly just be labeled by his listed position. The last thing he did was grab lineup data from stats.NBA.com and calculated positions using the initial set of clustering positions for lineup ordering. The results he found were solid despite the limited data, only using nearest defender data, which does not accurately reflect who's guarding whom for most of a possession. Overall, this article will be useful because it's just another strategy to think about when coming up with a cluster for my thesis. [20]

In this article Nicholas, uses NBA data to determine various position groups among players. Using k-means clustering he assigns a number at random from 1 to K to each of the observations (there were initial cluster assignments for each observation). Then he iterates until the cluster assignments stop changing. After that he then computes the cluster centroid for each k-cluster. The last thing he does is then assign each observation to the cluster whose centroid is closest, by calculating the cluster variation using the sum of the Euclidean distance between the data points and centroids. Later on, in the article he shows other ways on how to find the proper number of K's by using the elbow method and looking at his "gap" statistic. He then makes some very nice visualizations of the clusters and of the step-by-step methodology approach he took. He then breaks down each of his cluster groups, in this case he had eight different groups, which included high level starter fringe all-star, superstar all NBA, off the bench scoring 6th man, traditional big rebounder inside scoring, 3 and d, back to the basket scoring big man rebounder, all around role player, and another group which was an outlier so he couldn't name it anything specific. Overall, this article will be super helpful to look back at when coming up with my methodology, he does a very nice job of what he does step by step with some very cool visualizations on top of it. [10]

In this article Charles talks about clustering FIFA players. He says that he broke it down by four groups, which were goalkeepers, defenders, midfielders, and strikers. He then went in depth about his exploratory analysis. He uses this thing called AuDas to figure out how many clusters to use from the dataset. He tried looking at 11 clusters since those are the number of positions on the pitch. In conclusion he did end up producing 11 clusters with using AuDas. He said that there were only around 2 clusters that seemed a bit blurry, however, the rest of the clusters produced. He also concluded that the distinction was no longer on the roles but more on the quality of the players. Overall, I think this article will help me with clustering for sure, since I have a bunch of time learning different ways and programs to cluster things could be very useful for my project. [9]

In this article, the four authors, Evan Green, Michael Menz, Luke Benz, Gabriel Zanuttini-Frank, and Michael Bogaty talk about their way of clustering NBA players. They more or less say that they were defining types of NBA players based off how they play and how good they are. They then show a table of what player types they got, the trait that was overrepresented, the traits that were underrepresented, and sample players. In order to cluster they decided to use hierarchical clustering over k-means because they didn't like the random nature of k-means clustering and wanted stable results. They then wanted to look at what lineups from the clusters would be the nest. In order to do this, they looked at the lineup's total BPM. In conclusion

though, they felt that player types can't really predict a team's success, they said that they found no evidence that these player types can help predict lineup success better than simply the talent of the lineup. They suspect that this is because coaches and general managers implicitly understand and account for these player types. This article will definitely help with my thesis because it just shows me another way, I could cluster NFL players, especially since they used hierarchical clustering which a lot of the previous articles I've read didn't. [18]

This article by datanovia.com goes over how to run a Hopkins test in R-studio. The article talks about what code you need to use in order to run a Hopkins test in R. Some of this included what packages to use such as factoextra and clustertend. The article then goes further to explain how a Hopkins test works and how and when to accept or reject if a dataset is clusterable or not. This is referenced later on in this paper, and it will go in depth how the Hopkins test works. [14]

Within this article, the details of how hierarchical clustering is done and what applications it can be used for. Throughout the article the author, Doruk Kilitcioglu, gives definitions of what hierarchical clustering is and how it is used. These definitions are given throughout my thesis and will be referenced. The article also goes in depth of what type of dendrograms you can make when you use hierarchical clustering. [23]

In this lecture document by Cosma Shalizi, a very in-depth way of how ward's method of hierarchical clustering is done. He goes in depth about the different equations involved in the process and even gives examples of some dendrograms that result from some hierarchical clusters. This document is referenced within the thesis since it uses the formulas and definitions of wards method in it. Overall, this document does a very good job of summarizing and explain how hierarchical clustering works. [32]

In this article a lesson from datanovia goes over how to code and find the optimal number of clusters for a clustering analysis. This article goes over three different methods of finding the number of optimal clusters. The three methods are the elbow, silhouette, and gap-statistic methods. The article goes in depth about each of the different methods and how they are used. The article continues by showing how to run these methods in R, which is what this thesis used to find the number of optimal clusters. This article will be referenced throughout this thesis due to the definitions of the different methods the article provides. [16]

B. Statistics

In this article Bill discusses what statistics are most important in the game of football, and then goes in depth of what statistics are most important by position. Starting with team he says that DVOA (Defense – adjusted Value Over Average) and point differential are most important. For Quarterbacks he says that yards per attempt and adjusted net yards per attempt are the most important simple statistics, and that QBR is the best when comparing quarterbacks. Then moving to running backs he says that success rate from football outsiders is the most important stat to look at because it measures the rate at which a rusher keeps his offense on schedule. For wide receivers and tight ends, he says that catch rate, air yards per target, and receptions per route run are most important to look at. He then talks about what statistics are best for pass rushers which are quarterback knockdowns and sacks per knockdown. He then goes into special teams'

statistics which I'm not particularly interested in for my project, but it was still an interesting read. Overall, this article will be very useful because it will help me decide which variables I want to use for each position when clustering. [5]

This article talks about what trait/stat is the best for analyzing a position. For the quarterback, Andy says that the most important trait is accuracy and pocket movement. He says that Tom Brady has the trait while Blaine Gabbert does not. He then goes onto running backs which he thinks they need lateral agility to set up blocks and make defenders miss. He says that Le'Veon Bell has it while Darren McFadden doesn't. He then goes onto the perimeter receiver which needs to be good at beating press coverage, which he says Odell Beckham has, while Kelvin Benjamin does not. He then goes onto slot receiver. He says that they have to be great at changing direction to beat those linebackers, he says that Wes Welker has it but Jordan Mathews does not. He then goes onto the receiving tight end, which needs to have the trait of ball-tracking and being able to catch the ball at weird angles. He says that Antonio Gates has it while Maxx Williams doesn't. He then moves on to blocking tight ends, which he says they need to have contact balance which in other words means to have the ability to land blocks on the move. He says that Tyler Higbee has the trait while Eric Ebron does not. He then moves on to the offensive line, which he says that the most important trait for them is knee bending ability. This is because they get more leverage like that and "low man always wins" in football. He says that Tyron Smith has it while Eric Flowers doesn't. He then moves on to defensive end. He says that they must have edge-bending ability meaning they have to be flexible in parts of the body where others usually aren't. He says that Von Miller has it while Bjorn Warner doesn't. Then he talks about linebackers, which he says that the most important trait for them is play recognition. He says that Luke Kuechly has it while Darron Lee does not. Then he talks about cornerbacks, and he thinks that cornerbacks need to have the man-to-man coverage ability. He says that Jalen Ramsey has it while Dee Milner does not. The last position he talks about is Safety, where he says that safetys must be great at open field tackling. He says that Earl Thomas has it while Matt Elam does not. Overall, this article will be helpful for my thesis because it will help me choose which stats, I want to use for each position based off these traits. [7]

In this article, Allen talks about what he thinks is the "most important stat" in football. He says that Bud Goode, the father of pro football analysis, told the world about the most important stat which is net yards per attempt. He goes in depth in the article about how Goode worked for multiple NFL franchises, with a lot of them being very successful using his algorithm. He talks about how it's calculated which is (passing yards – sack yards) divided by (passes attempted + times sacked). He says that if you go to the box score of the game and look at who had the better NY/A, 80% of the time, the team with the better NY/A wins. He then goes on to say that the second most important stat to look at is interception percentage. Then he goes on to talks more history about NY/A and how passing is the best way of telling how a team will win. He goes on in the article to say that average yards per rush is no measure of success and that establishing the run isn't key. Overall, I think this article will help because it gives me stats to consider putting more weight on, and stats to put a lot less weight on. In fact, I thought setting up the run was a huge thing but according to this guy it's not, so it's really just giving me more perspectives and ways to think about what stats I should consider using for my thesis. [6]

In this article, Scott gives a great breakdown of some of the stats that are super important to understand that most fans do not. The first stat he begins to talk about is passer rating which he says everyone loves to talk about, but no one actually loves. He says that people usually get confused by passer rating and quarterback rating due to a lot of people looking at ESPN when looking up statistics. Scott says that passer rating simply measures the passes a quarterback attempts without being sacked in a game, and the math is built around decade-old averages that probably could use an update as the stat is useless to compare quarterbacks across eras. The next stat he talks about is time of possession. He says that of all the stats you can look at from a simple box score after the game, one of the last to pay attention to is time of possession. He really makes a good point about it when he says the actual possession of the ball is not important if you are not scoring any points. He then goes on to say that a long drive that ends in a punt or turnover looks good for your time of possession, but what did it actually do to your win probability? It's a great point and I definitely respect his opinion. Then next stat he talks about is rushing yards per carry which he says one of the stats that should be a good indicator of effectiveness is rushing yards per carry (YPC). He goes on to say that the problem is most runs in the NFL are gains of two to four yards, and the YPC number is inflated by those rare, long runs. A long run is great for an offense, but it can only help you on that one drive. He says that the real stat to look at should be carries. He then goes on to talk about the stats of preseason stats, fourth quarter comebacks, and punting average. Overall, I think this article will be super helpful when deciding which stats, I'm going to use for the clusters. This will help me weave out the important and non-important stats. [21]

In this article John talks about what stats are most important when looking at wins. The first thing he starts about talking about is turnovers. He says that when he broke it down by year, it looked like: Eight top-10 teams in takeaways finished with winning records in each season between 2012-14. In 2013, there were 12 teams tied in the top 10 and last year there were 13. So, he says that in an individual season it matters a lot more than, say, over a three-year stretch. The next stat he talks about is rushing yards per game. He says that six of the top ten teams in rushing yards per game posted winning records in 2014. The next stat he says is super important is sacks, where he says that five of the top 10 teams posted winning records, but the top two teams did not. He then brings up his next statistic, which is passing yards allowed per game, where he says seven of the top 10 teams posted winning records. The next stat he talks about after that is yards allowed per game where he says nine of the top 13 teams had winning records. The last stat he talks about is points allowed per game. He says that this stat is the most important and revealing stat when talking about defense. He gives a statistic where eight of the top 10 teams when looking at the past three years from when the article was written have winning records. Overall, this article will be helpful because it will continue to give me ideas of what stats I should use when coming up with my clusters. [22]

In this article Bill talks about the five most important football statistics according to him. The first stat he talks about is explosiveness. He breaks this stat down by the yards per play margin and the average scoring margin. The next one he talks about is efficiency. In order to observe this, he looked at the success rate margin not including garbage time. He got this stat from football outsiders. The next stat he looked at was field position. In order to look at this he looked at a field position margin rate, in order to see how many times, they had a certain field position and then looked at the certain amount of time the team was there that game. The next

statistic he looks at is finishing drives. To do this he defined scoring opportunities are when you are inside the opposing team's 40. He then looked at the team's average points when inside the 40. The last stat he looks at is turnover margins, which is obviously a huge determiner of the winner of the game. Overall, I think this article will be super helpful for my thesis because it will help me weed out which statistics I should choose for the clusters and lineups. [11]

This article by Nick discusses the best offensive lines in the NFL when looking at expected rushing yards. In the beginning of the article, he gives a small recap of some of the running backs that have had the best expected rushing yards and such. He then gets into what the article really tries to convey. He provides a glossary of the statistics he is going to talk about in the article. Some of these stats include yards per carry, expected yards per carry, expected rushing yards, rushing yards over expectation, and lastly rushing yards over expectation per attempt. The first team he shows is the Ravens who had the most rushing yards and expected rushing yards. He then moves to show the second-best offensive line team, which were the Eagles who had a little less of expected rush yards. The next team was the Cardinals, there expected yards per carry was 4.5 which was .12 less than the Eagles which is kind of significant an interesting to look at how different even the top 3 teams were in this stat. The next two teams were the Falcons and then lastly the 49ers. This article will be super helpful for my thesis because it will help me formulate ideas of how I can come up with stats that would differentiate offensive linemen. This will also give me a good understanding of what one person views as being a good offensive lineman. [33]

In this article Jeff talks about defensive pass statistics. I looked into this article because I was trying to decipher how I'd classify zone versus man corner backs during my analysis. I think this article could help me find a way of figuring this out. The first stat he talks about is opponent receptions plus/minus which is basically the number of incompletions over expected that a team creates. He then used the pass coverage classification system to figure out if they were playing man or zone. He first looks at players in man coverage. He goes on to say that last year Stephon Gilmore led the league in this stat and there was actually a linebacker in the top 5 in Eric Kendricks which I did not expect. He also says that this stat is mostly looking at man to man coverage which is good to know if I am using this stat for my thesis. He then uses the pass coverage classification system again to look at players in zone coverage while looking at the same stat of opponent receptions plus/minus. In this case Denzel Ward topped the list. He said that this was less convincing than other measurements he had in the first table. He says that some of the top players in the table weren't that great during the season to say the least. He concludes the article saying that this stat might not be the best to look at when predicting a corners play from season to season, however, this is still a very interesting stat and something I would want to look more into. [15]

In this article Jeff talks about how he thinks that receivers are the most important position in football, replacing the Quarterback. This article interested me because my goal is to when after I finish the player classification part, I am going to try and create an optimal lineup of the player styles, and in order to do this I would need to put different weights on the different positions. In order to do this, he uses the offensive share metric to see how valuable a player is. When looking at the database of OSM's he said that you will notice that there are a lot of pass catchers with a higher OSM than the passers. He says that this is because a receiver's ability is

dependent on the individual compared to the other positions that rely on other's more. He then shows a chart that the gap between the receivers and quarterbacks OSM's as time goes on. He goes on to say that the value of quarterbacks is declining because they have more time to throw and because of this are throwing safer passes. He says that receiver values are increasing because the data says receivers are gaining more yards of separation at the time of the throw while also obtaining more yards after the catch than expected. He goes on in the article showing different charts of the value changing for receivers and quarterbacks' overtime. He then concludes the article by saying that receivers just might be more valuable than quarterbacks nowadays because he says that OSM certainly provides us the context of value in terms of individual behavior, what it lacks is value in terms of the environment. He then says that OVS might be a better indicator which is something I could definitely look into. [28]

This github was used for the nflscrapR data. This github repository provided by Ron Yurko contained the player-season statistics that were used for the clustering analysis. The data was from 2009-2017 and had the statistics needed in order to comply the quarterback, running back, and wider receiver datasets. The repository has more than just player-season statistics however, it also includes by game statistics and more. This repository will be referenced within the thesis since the data used came from it. [30]

This htm webpage was used for the winning percentage data used in the thesis. The data collected was from the years 2009-2017, and the variables collected were the winning percentages of all thirty-two NFL teams throughout those seasons. This was used in the thesis in order to compare clusters of the same position dataset using the winning percentage statistic. This will be referenced in the thesis since some of the data used in it is from this dataset. [1]

C. QB Clustering

This article uses the mathematical method of k-means clustering in order to start classifying Quarterbacks. Each cluster can be described by the mean values of each variable thereby allowing us to compare and contrast clusters and thereby characterize the quarterbacks within them. After deciding to use clustering techniques, they then needed to decide what variables they wanted to subset. They decided to use 9 variables some of which came from their own website. Surprisingly, yards per attempt is a big variable to use when clustering Quarterbacks. After clustering, they ended up getting 6 different group of clusters. Cluster 1 is the cluster of truly elite quarterbacks, while Clusters 2 and 3 were players that were second tier in a given season, Clusters 4 and 5 were third-tier guys and Cluster 6 were the quarterbacks who played poorly that year. Clusters 2 and 4 were of the safer variety, while 3 and 5 were riskier players. In conclusion, the article states that quarterback clustering has allowed us to improve the way we talk and evaluate quarterback play. [12]

D. WR Clustering

In this article Steven Morse came up with an idea to look at Cole Beasley statistics in order to verify if Dak Prescott's statement of Cole being able to "stretch defenses". In order to do this, he had to scrape individual player outcomes for all active receivers from pro football reference. Then he wanted to see if there were players with mostly short plays but a considerable rate of "huge plays". Then the last thing he wanted to do was cluster similar players based on their entire distribution of catch yardage rates. After the cluster was made, he was able to make 7

different groups of wide receivers to compare. When he was all finished with his clusters, he concluded that although Cole Beasley was considered a "check down king" according to his cluster, he said that it didn't necessarily mean Dak was wrong in what he said. Some other thoughts he had was to look at the distribution of yards per game, instead of yards in different amounts, as well as maybe using and going into expected points added. [27]

In this article Cory starts off by telling us where he got his data from, adavancedfootballanalytics.com, because he needed a stat called WPA, which stands for Win Probability Added. The site says that WPA is important because, 'although we still can't separate an individual player's performance from that of his teammates, we add up the total WPA for plays in which individual players took part. This can help us see who really made the difference when it matters most'. He then went on to show which variables he decided to use, which includes 14 different statistics. He then started clustering by using hierarchical clustering. In the end he was able to break up the clusters into 5 different groups, in order to compare the different types of wide receivers. The last thing he did was compare the WPA of the different clusters to show how different in skill each wide receiver group is. [26]

In this article Ridley talks about and shows how to find the next Julio Jones. In order to do this, he looked at veteran and young receivers at the start of their careers. The first thing he looked at was their combine stats, which were the forty, weight, height, vertical, arm length, draft position, bench, broad, shuttle, and cone. The next thing he looked at were in game stats, which were yards per season average, pass attempt percentage of WR group, pass percentage by location, yards per completion, points scored, completion percentage, and down and distance percentage. He then used a cluster a hierarchical cluster method, specifically an agglomerative cluster which he says it can handle a large feature space, he can define his own number of clusters, and its flexible in linkage and distance metrics he chose. There's a link to where he does this analysis and it's very cool thing. Overall, this article will help me formulate new ways of clustering, and even shows me some new statistics that I didn't really think about using. [25]

E. Offensive Line Clustering

In this article Sean talks about how he tries to develop an empirical method to value offensive lineman in the NFL and then compare his results with their relative salaries. He makes a point that there are no true output statistics specific to an offensive lineman's performance and that every statistic that could be relevant is an empirical measure of someone else's performance, like the RB or QB. He says that while this argument of shared statistics could be made with the QB throwing to a great WR, ultimately there is always the confounding effects of a teammate on an individual player's performance, and that the fact remains that there are no stats that even point to an offensive lineman's performance. Since this is the case, he has to use statistics for different situations. For running performance, he uses power success percentage, rb yards, stuffed percentage, second level yards, and open field yards. He then grouped 116 offensive lineman into 8 clusters using k-means and found that there are definitely linemen that are under and overvalued. This can definitely help me with my project because it's giving me a perspective of how to cluster o-line. [29]

Data Summary:

There were three datasets obtained throughout this process. The first data set put together was scraped from "pro-football-reference.com" [1]. This data included variables such as team, season, wins, losses, win percentages, point scored, etc. The data used from this however, was just the team, season, and win percentage of NFL teams from the 2009-2017 seasons. This made the final dataset 288 observations with 3 variables.

The next two datasets scraped were from the nflscrapR GitHub [30]. The first dataset scraped from this was the roster dataset. The roster data includes season, player name, team, position, and GSIS_ID from the 2009-2017 seasons. With this data all that was needed was the position, GSIS_ID, and team of the given player. GSIS_ID is just the ID given to players in the datasets from nflgsis.com.

After this was done to the dataset, the roster data was 2,321 observations by 3 variables. The last dataset retrieved was the season player statistics data from the nflscrapR GitHub. This included passing, rushing, and receiving data from the 2009-2017 NFL seasons. This data didn't have the position or team of the observed player which meant I had to merge the roster dataset and the season player statistics data by the player's GSIS_ID. Once this was done, the data was broken into three position groups, Quarterback, Running back, and Wide Receiver.

Each of these datasets had passing, rushing, and receiving variables, however, some variables differed based on the position in the data set. Each of the datasets had a column created called either Receiver ID, Passer ID, or Rusher ID, which was comprised of a player's GSIS_ID,

Variable	Description
Passer_ID	The GSIS_ID, Player Name, and Season separated by dash marks.
Attempts	number of pass attempts
Completions	number of completed passes
qb_drives	drives as passer
Comp_Perc	completion percentage
passing_yards	total yards gained from passing
total_raw_airyards_qb	total air yards thrown (both complete and incomplete passes)
Total_Comp_AirYards	total air yards from completions
Yards_per_Att	yards gained from passing per attempt
Yards_per_Comp	yards gained from passing per completion
pass_yards_per_drive	yards gained from passing per drive
Raw_AirYards_per_Att	air yards thrown per pass attempt (all pass attempts)
Comp_AirYards_per_Att	air yards thrown per pass attempt (only completed passes)
Comp_AirYards_per_Comp	air yards thrown per completion (only completed passes)
Raw_AirYards_per_Drive	air yards thrown per drive (all pass attempts)
Comp_AirYards_per_Drive	air yards thrown per drive (only completed passes)
TimesHit	number of times the QB was knocked down (includes being hit on pass attempts)
TimesHit_per_Drive	number of times the QB was knocked down (includes being hit on pass attempts) per drive
Interceptions	number of interceptions
passing_tds	number of touchdowns thrown
TD_per_Att	TDs thrown per pass attempt
Int_per_Att	INTs thrown per pass attempt
TD_per_Comp	TDs thrown per completion
TD_per_Drive	TDs thrown per drive
Int_per_Drive	INTs thrown per drive
total_clutch_epa_qb	total EPA from pass attempts weighted by each play's WPA
Targets	number of targets
Carries	number of carries
rushing_yards	total yards gained from rushing
fumbles	number of fumbles
rushing_tds	number of TD rushes
EPA_qb	total win probability added from pass attempts
WPA_qb	total expected points added from pass attempts

Table 1: *Brief Description of variables used for clustering Quarterbacks*

Variable	Description
Rusher_ID	The GSIS_ID, Player Name, and Season separated by dash marks.
Carries	number of carries
hb_drives	drives as ball carrier
Car_per_Drive	carries per drive
rushing_yards	total yards gained from rushing
Yards_per_Car	yards gained per carry
rush_yards_per_drive	rushing yards per drive
fumbles	number of fumbles
rushing_tds	number of TD rushes
TD_per_Car	TDs per carry
Fumbles_per_Car	fumbles per carry
fumbles_per_drive_hb	fumbles per drive
TD_Drive	TDs per drive
total_clutch_epa_hb	total EPA from carries weighted by each play's WPA
Targets	number of targets
Receptions	number of receptions
Targets_per_Drive	number of targets per drive
Rec_per_Drive	number of receptions per drive
rec_yards	total yards gained from receiving
rec_yards_per_drive	total yards gained from receiving per drive
Total_Raw_YAC	total of yards after catch (YAC) as well as yards missed from incomplete passes
Rec_Percentage	reception percentage
rec_tds	number of TD receptions
qb_drives	drives as passer
Attempts	number of pass attempts
pass_yards	total yards gained from passing
pass_tds	number of touchdowns thrown
EPA_hb	total expected points added from all carries
WPA_hb	total win probability added from all carries

Table 2: *Brief Description of variables used for clustering Running Backs*

Name, and Season. This way when the data is filtered by having the ID column unique for each of the datasets, it results into having player-season statistics provided by the data for all the Quarterbacks, Running backs, and Wide Receivers from the 2009-2017 NFL seasons. The final dimensions of the Quarterback dataset were 645 observations by 33 variables, and the variables used can be seen in Table 1. The final dimensions of the Running back dataset were 1,208 observations by 29 variables, and the variables used can be seen in Table 2. The final dimensions of the Wide Receiver dataset were 1,800 observations by 30 variables, and the variables used can be seen in Table 3.

Variable	Description
Receiver_ID	The GSIS_ID, Player Name, and Season separated by dash marks.
Targets	number of targets
Receptions	number of receptions
rec_drives	drives as an eligible receiver
Targets_per_Drive	number of targets per drive
Rec_per_Drive	number of receptions per drive
receiving_yards	total yards gained from receiving
receiving_yards_per_drive	total yards gained from receiving per drive
Total_Raw_YAC	total of yards after catch (YAC) as well as yards missed from incomplete passes
Yards_per_Target	yards gained per target
YAC_per_Target	yards gained after catch per target
Total_Caught_YAC	total yards gained after catch for receptions
Total_Dropped_YAC	total yards missed from incomplete passes
Dropped_YAC_per_Target	yards missed after drop per target
YAC_per_Rec	yards gained after catch per reception
YAC_per_Drive	yards gained after catch per drive
Rec_Percentage	reception percentage
receiving_tds	number of TD receptions
TDs_per_Drive	TDs caught per drive
TD_per_Targets	TDs caught per target
TD_per_Rec	TDs caught per reception
clutch_epa_rec	total EPA from targets weighted by each play's WPA
Attempts	number of pass attempts
passing_yards	total yards gained from passing
passing_tds	number of touchdowns thrown
Carries	number of carries
rushing_yards	total yards gained from rushing
rushing_tds	number of TD rushes
EPA_rec	total expected points added from all targets
WPA_rec	total win probability added from all targets

Table 3: *Brief Description of variables used for clustering Wide Receivers*

Once the datasets were put together, correlation plots for each of the datasets were made. This was done in order to see what variables correlated with each other and in particular the WPA of each of the datasets. Doing this shows what variables correlate most with raising your team's chance to win

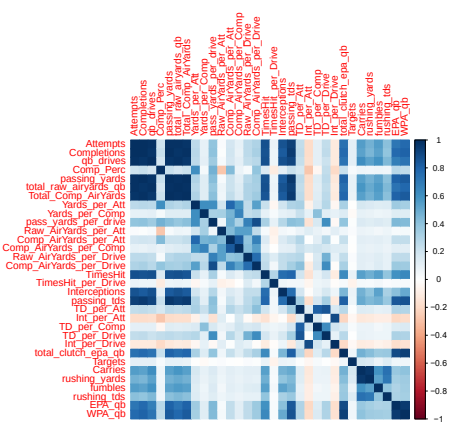


Figure 1: *Correlation of variables used from Quarterback Dataset*

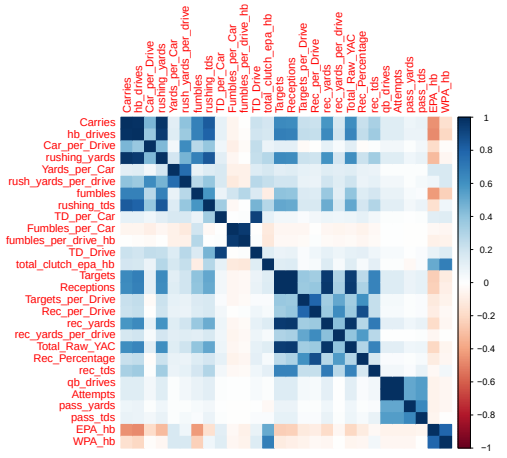


Figure 2: *Correlation of variables used from Running Back Dataset*

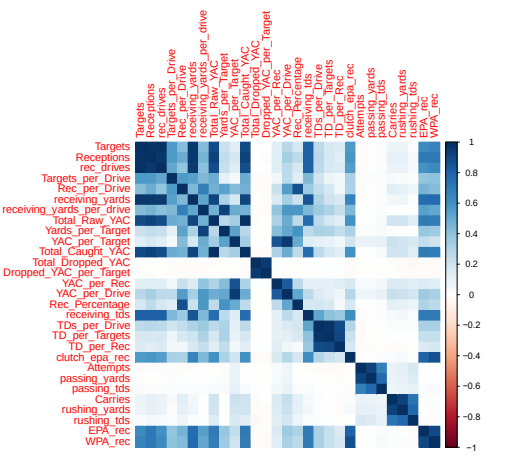


Figure 3: *Correlation of variables used from Wide Receiver Dataset*

for the given position. Figure 1 is for the Quarterback dataset, Figure 2 is for the Running back dataset, and Figure 3 is for the Wide Receiver dataset.

When observing Figure 1, WPA_qb had multiple variables highly correlated with it. These variables were attempts, completions, passing_yards, passing_tds, and total_clutch_epa_qb. Later in the analysis when comparing quarterback clusters, these are some of the statistics that will be analyzed.

While dissecting Figure 2, it can be seen that there are only a couple variables that are correlated with WPA_hb. These variables are yards_per_car, rush_yards_per_drive, TD_per_Car, and total_clutch_epa_hb. When comparing running back clusters later in the analysis, some of these statistics mentioned will be examined.

When inspecting Figure 3, receiving_tds, clutch_epa_rec, receiving_yards, targets, and receptions are the variables most correlated with WPA_rec. Some of these variables will be used to compare wide receiver clusters later on in the analysis.

Methodology:

In order to figure out whether or not the datasets are fit to be clustered I conducted a Hopkins test. The *Hopkins statistic* is used to assess the clustering tendency of a data set by measuring the probability that a given data set is generated by a uniform data distribution. In other words, it tests the spatial randomness of the data [14].

For example, let D be a real data set. The Hopkins statistic can be calculated as follow:

1. Sample uniformly n points (p_1,…, p_n) from D.
2. Compute the distance, x_i, from each real point to each nearest neighbor: For each point $p_i \in D$, find it's nearest neighbor p_j; then compute the distance between p_i and p_j and denote it as $x_i = dist(p_i, p_j)$
3. Generate a simulated data set ($random_D$) drawn from a random uniform distribution with n points (q_1,…, q_n) and the same variation as the original real data set D.
4. Compute the distance, y_i from each artificial point to the nearest real data point: For each point $q_i \in random_D$, find it's nearest neighbor q_j in D; then compute the distance between q_i and q_j and denote it $y_i = dist(q_i, q_j)$
5. Calculate the Hopkins statistic (H) as the mean nearest neighbor distance in the random data set divided by the sum of the mean nearest neighbor distances in the real and across the simulated data set [14].

The formula is defined as:

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

The Null hypothesis is that the data set D is uniformly distributed (i.e., no meaningful clusters) and the alternative hypothesis is that the data set D is not uniformly distributed (i.e., contains meaningful clusters) [14].

We can conduct the Hopkins Statistic test iteratively, using 0.5 as the threshold to reject the alternative hypothesis. That is, if H < 0.5, then it is unlikely that D has statistically significant clusters. Put in other words, If the value of the Hopkins statistic is close to 1, then we can reject the null hypothesis and conclude that the dataset D is significantly a clusterable data [14].

When running the Hopkins test on the three datasets the value of H is as follows. For the Quarterbacks dataset H was 0.9004466 which means we can reject the null hypothesis in favor for the alternative that this dataset has statistically significant clusters. For the Wide Receiver dataset H was 0.970285 which means we can again reject the null hypothesis in favor for the alternative that this dataset has statistically significant clusters. For the Running back dataset H was 0.9409178 which means we can again reject the null hypothesis in favor for the alternative that this dataset has statistically significant clusters.

Now that it is known that all three of the datasets have statistically significant clusters, I had to decide on what type of clustering technique to use. In this case I decided to go with hierarchical clustering. This is due to the fact that Hierarchical clustering is a powerful technique that allows you to build tree structures from data similarities. With it you can see how different sub-clusters relate to each other, and how far apart data points are [23]. Since this research is trying to group players based on their statistics, this clustering method seemed to be the best to work with.

The algorithm:

1. Start with each point in a cluster of its own
2. Until there is only one cluster
    (a) Find the closest pair of clusters
    (b) Merge them
3. Return the tree of cluster-mergers

Then within hierarchical clustering there are different methods in which you can cluster. In this research specifically Ward's method was used. Ward's method says that the distance between two clusters, *A* and *B*, is how much the sum of squares will increase when we merge them [32]:

$$\begin{aligned}\Delta(A,B) &= \sum_{i\in A\cup B}\|\vec{x}_i-\vec{m}_{A\cup B}\|^2-\sum_{i\in A}\|\vec{x}_i-\vec{m}_A\|^2-\sum_{i\in B}\|\vec{x}_i-\vec{m}_B\|^2\\ &= \frac{n_An_B}{n_A+n_B}\|\vec{m}_A-\vec{m}_B\|^2\end{aligned}$$

Where m_j is the center of cluster *j*, and n_i is the number of points in it. Δ is called the merging cost of combining the clusters *A* and *B*. With hierarchical clustering, the sum of squares starts out at zero (because every point is in its own cluster) and then grows as we merge clusters. Ward's method keeps this growth as small as possible. Notice that the number of points shows up in Δ, as well as their geometric separation. Given two pairs of clusters whose centers are equally far apart, Ward's method will prefer to merge the smaller ones [32].

Once hierarchical clustering using the ward method was decided to be used as the clustering technique, the next step was to decide how many clusters for each dataset should be used. There are three statistical methods for determining the optimal number of clusters used, which are the elbow method, average silhouette method, and the gap-statistic method.

The elbow method looks at the total WSS (total within-cluster sum of square) as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't improve much better the total WSS. The optimal number of clusters can be defined:

1. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
2. For each k, calculate the total within-cluster sum of square (wss).
3. Plot the curve of wss according to the number of clusters k.
4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters. [16]

The average silhouette approach measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering. Average silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values for k [16].The algorithm is similar to the elbow method and can be computed as follows:

1. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
2. For each k, calculate the average silhouette of observations (*avg.sil*).
3. Plot the curve of *avg.sil* according to the number of clusters k.
4. The location of the maximum is considered as the appropriate number of clusters. [16]

The gap statistic method compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data. The estimate of the optimal clusters will be value that maximize the gap statistic (i.e, that yields the largest gap statistic). This means that the clustering structure is far away from the random uniform distribution of points. The algorithm works as follows:

1. Cluster the observed data, varying the number of clusters from k = 1, …, k_{max}, and compute the corresponding total within intra-cluster variation W_k.
2. Generate B reference data sets with a random uniform distribution. Cluster each of these reference data sets with varying number of clusters k = 1, …, k_{max}, and compute the corresponding total within intra-cluster variation W_{kb}.
3. Compute the estimated gap statistic as the deviation of the observed W_k value from its expected value W_{kb} under the null hypothesis: $Gap(k) = 1/B\ \Sigma^{B}_{b=1}\ log\ (W^{*}_{kb})$ - $log\ (W_K)$. Compute also the standard deviation of statistics.
4. Choose the number of clusters as the smallest value of k such that the gap statistic is within one standard deviation of the gap at k+1: $Gap\ (k) \geq Gap\ (k + 1) - s_{k+1}$. [16]

The results of these methods on each of these datasets can be seen in the appendix. Looking at all three methods for the Receiver dataset, which can be seen in Figures 4, 5, and 6, I came to the conclusion that the optimal number of clusters should be 5, given that the silhouette and gap statistic method said it should be 2 clusters, while the elbow method had a bend at 2 and 6 clusters. For thew purpose of this research however I decided more clusters would be better to analyze than less and so I decided to go in between and choose 5 clusters.

Looking at all three methods for the Running Back dataset, which can be seen in Figures 7, 8, and 9, I came to the conclusion that the optimal number of clusters should also be 5, given that the silhouette method suggested 2 clusters, the gap statistic method suggested 9 clusters, and the elbow method had a bend at 2 and 5 clusters. Since the optimal number of clusters varies from 2, 5, and 9, I decided to go in the middle and choose 5 clusters for the Running back dataset as well.

Lastly, looking at all three methods for the Quarterback dataset, which can be seen in Figures 10, 11, and 12, I came to the conclusion that the optimal number of clusters again, should be 5, given that the silhouette method suggests 3 clusters, the gap statistic method suggests 9 clusters, and the elbow method has a bend at 2 and 4 clusters. With the optimal clusters varying from 3, 4, and 9 clusters from the methods, it was decided again to make the optimal number of clusters 5 for this dataset, since that number is around the middle of the suggested number of clusters.

Clustering Results:

In Figures, 13, 14, and 15 are the distributions of the 5 clusters for each of the 3 datasets:

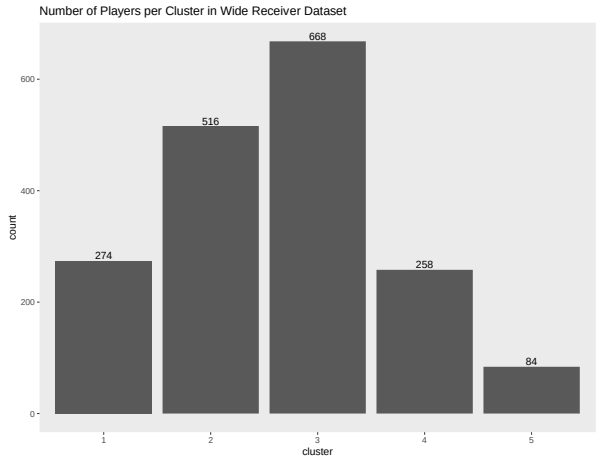


Figure 13: *Distribution of Wide Receivers among the five clusters*

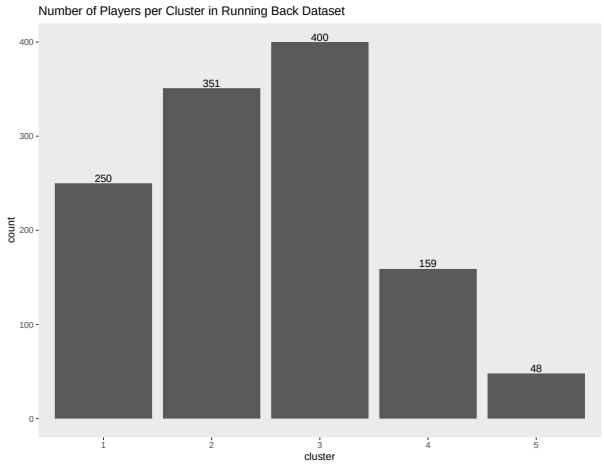
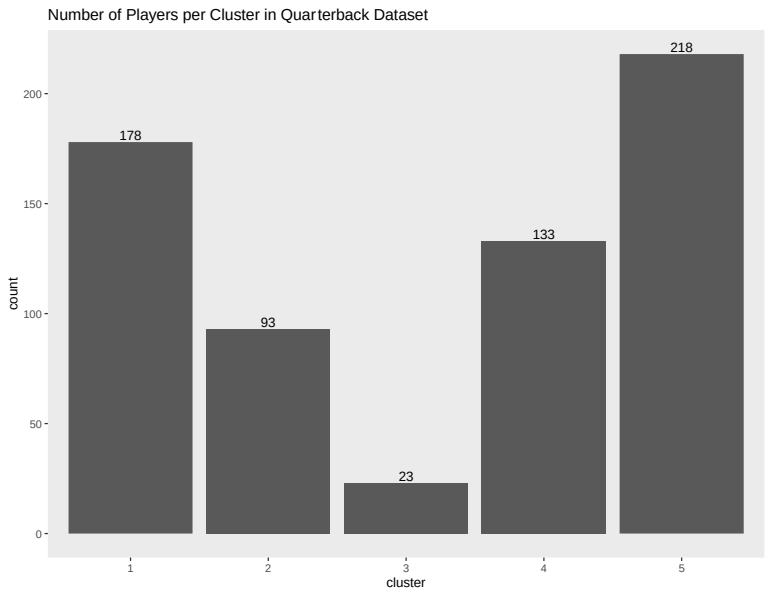


Figure 14: *Distribution of Running Backs among the five clusters*

Figure 15: *Distribution of Quarterbacks among the five clusters*



The wide receivers had a total of 1,800 player-season observations, which can be seen distributed into five different clusters in Figure 13. The Running backs had a toal of 1,208 player-season observations, which can be seen distributed into five different clusters in Figure 14. The Quarterbacks had a total of 645 player-season observations, which can be seen distributed into five different clusters in Figure 15.

As you can see for each of the positions there are some clusters that have a lot more players in one cluster or not that many in another. Once the cluster groups were made for the three different datasets, the cluster groups were then analyzed against the average WPA of their group to see which cluster group gives a team the best chance to win. The results can be seen in Figures 16, 17, and 18 below:

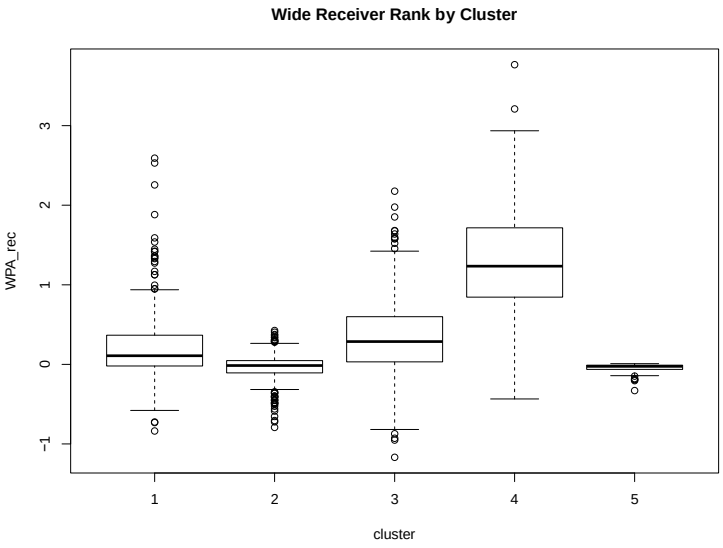


Figure 16: *Wide Receiver Clusters versus WPA boxplots*

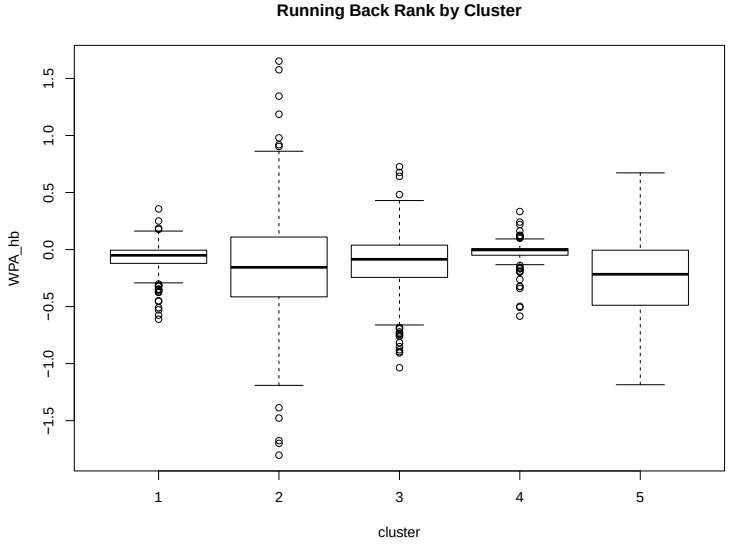


Figure 17: *Running Back Clusters versus WPA boxplots*

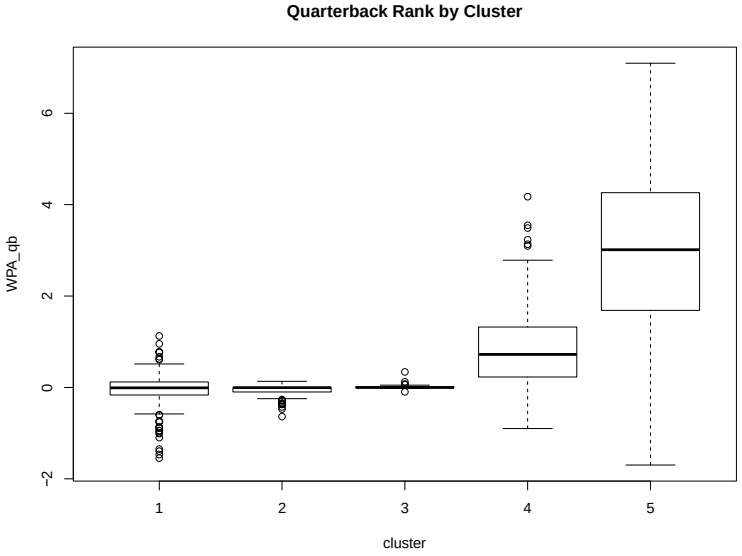


Figure 18: *Quarterback Clusters versus WPA boxplots*

In Figure 17 there is one Wide Receiver cluster group that is clearly greater than the rest. However, the rest of the clusters seem to all have around the same WPA. In Figure 17 all of the Running Back clusters are around the same WPA. In Figure 18 one of the Quarterback clusters is clearly greater than the rest. The rest of the clusters are then similar in WPA. The best cluster in terms of average WPA for the Wide Receivers was cluster number 4. The best cluster in terms of average WPA for the Running Backs was cluster number 4 as well. For Quarterbacks, the best cluster in terms of average WPA was cluster number 5.

Once the best clusters in terms of average WPA were determined, they were then analyzed to see what statistics were above or below average for their position. The clusters were broken into boxplots with the standardized data. The statistics chosen to observe were ones that were highly correlated with the WPA of the dataset, and other statistics that were important to help interpret the cluster based on the position of the dataset. Attempts was also added in order to see if any groups had players who play multiple positions, such as Taysom Hill. If a boxplot is green that means the mean of the statistic from the green boxplot observed from the certain cluster, is above the average for that statistic in the given position group. If a boxplot is red that means the mean of the statistic from the red boxplot observed from the certain cluster, is below the average for that statistic in the given position group. The dashed line is the standardized value of 0 meaning it's the average of the observed statistics. These clusters can be seen in Figures 19, 20, and 21 below. The rest of the clusters can be seen in the appendix:

Figure 19: *Breakdown of the best Wide Receiver cluster in terms of average WPA*

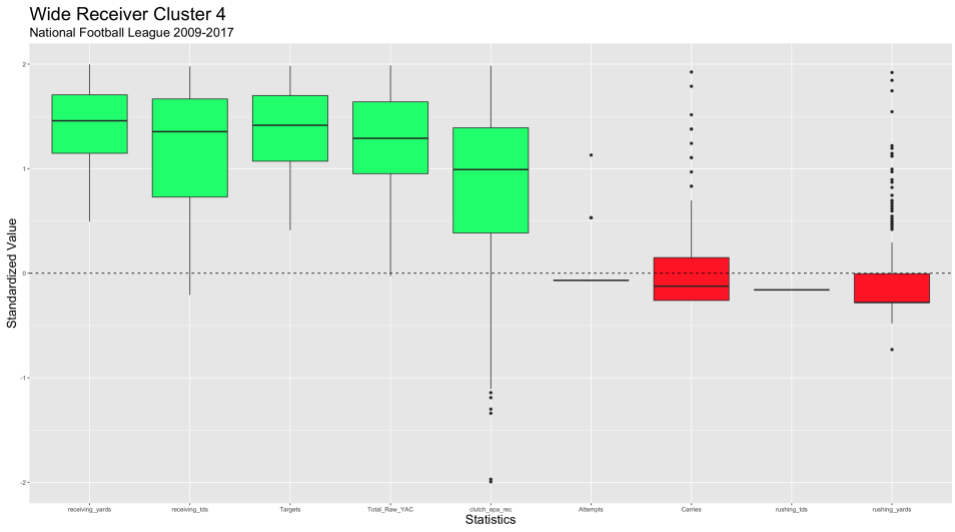


Figure 20: *Breakdown of the best Running Back cluster in terms of average WPA*

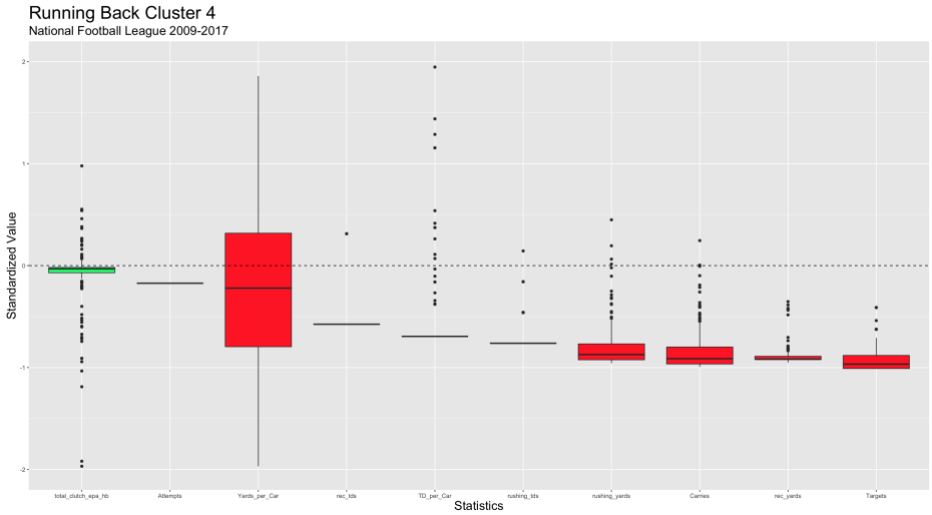
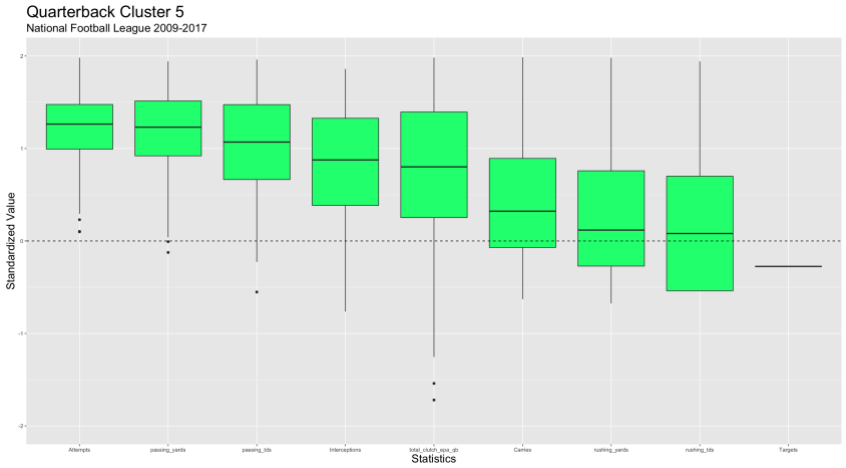


Figure 21: *Breakdown of the best Quarterback cluster in terms of average WPA*



Looking at Figure 19 you can clearly see why this cluster was the best in terms of average WPA for Wide Receivers. This group is way above the average in statistics such as receiving yards, receiving touchdowns, targets, total raw yards after catch, and clutch EPA, which are all stats you need to be good at if you want to help your team win as a wide receiver. The stat the cluster group had its lowest average in was rushing yards.

Looking at Figure 20 some might question as to how or why this cluster had the highest mean WPA. All of this cluster's statistics are below average in the observed stats besides clutch EPA. The statistic with the lowest average was targets.

Looking at Figure 21 you can observe that all of the statistics given were above average except for one. This cluster group makes sense to be the best cluster in mean average WPA since this group is largely above the average in statistics such as passing yards and passing touchdowns. The stat with the lowest average was target, meaning this group of quarterbacks strictly passes the ball.

After observing Figure 20 and wondering why that group had the highest average WPA, I decided to test the clusters against the actual average winning percentages of the clusters. In order to do this the data needed to be filtered out so that it only contained players who were on the same team the whole season of a given season. This had to be done in order to not get multiple winning percentages for same player-seasons in the dataset. This made the clustered Receiver dataset go from having 1,800 players to 812 players, the clustered Running Back dataset from having 1,208 players to 536 players, and the clustered Quarterback dataset from having 645 players to 264 players. Once the data was ready, the clusters were then compared to the average win percentage of the cluster group. This can be seen in the bar graphs below in Figures 34, 35, and 36:

Figure 34: *Breakdown of Wide Receiver clusters versus average Win Percentage*

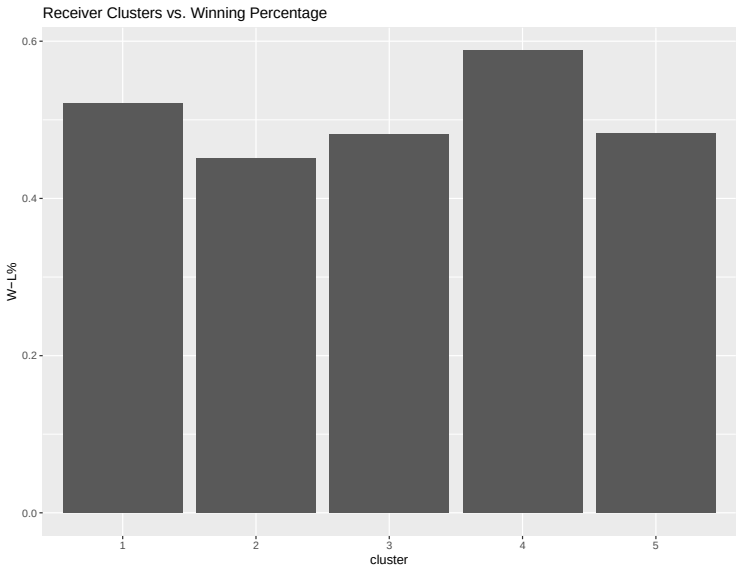


Figure 35: *Breakdown of Running Back clusters versus average Win Percentage*

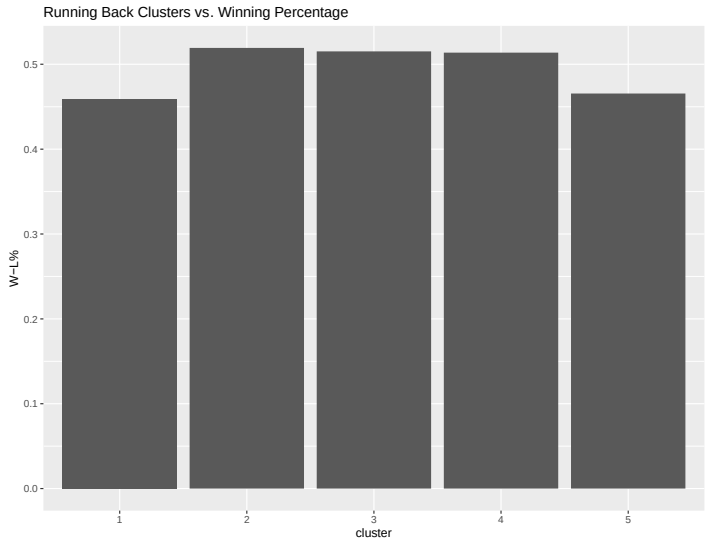
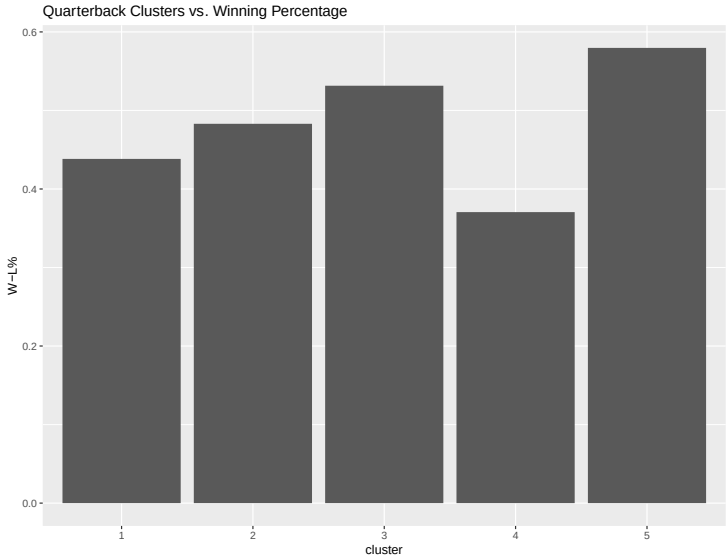


Figure 36: *Breakdown of Quarterback clusters versus average Win Percentage*



From observing Figure 34 you can see that cluster 4 is still best Wide Receiver cluster, in this case in terms of average winning percentage. When observing Figure 35 cluster 2 is the best Running Back cluster in terms of average winning percentage. This was different when this cluster was observed against WPA. However, when looking at cluster 2's breakdown in Figure 27, all of the statistics are above average except for attempts. This makes a lot more sense compared to Running Back's cluster 4 being the best in WPA when having relevant statistics all below average. Observing Figure 36 it is easy to tell that cluster 5 is still the best Quarterback Cluster, in terms of average winning percentage this time.

After observing the clusters against winning percentage, I wanted to see if I could define the clusters somehow. In order to formalize this, tables were created for each of the positions that shows, the cluster number, description, best relevant stats, worst relevant stats, best players in terms of WPA, worst players in terms of WPA, WPA rank, winning percentage rank, and numbers of players within in each cluster. The tables can be seen below in Tables 4, 5, and 6:

Table 4: *Description of Wide Receiver Clusters*

WR Cluster	Description	Best Relevant Stats	Worst Relevant Stats	Best Players (highest WPA)	Worst Players (lowest WPA)	WPA Rank (mean wpa)	Win Percentage Rank (mean winpct)	Number of Players
1	**Dual Purpose Receivers**	rushing_tds / receiving_tds	targets / total_raw_yac	'12 Golden Tate / '14 Antonio Brown	'14 Cordarrelle Patterson / '14 Cecil Shorts	3rd	2nd	274
2	**Low-Production Backup Receivers**	rushing_tds / carries	targets / receiving_yards	'09 Craig Davis / '14 James Wright	'17 Kamar Aiken / '12 Legedu Naanee	4th	5th	516
3	**Middle of the Field Threats**	targets / receiving_yards	receiving_tds / rushing_yards	'10 Braylon Edwards / '12 Brandon Gibson	'11 Mike Thomas / '12 Larry Fitzgerald	2nd	4th	668
4	**High-Usage Red-Zone Threats**	receiving_yards / receiving_tds	rushing_yards / rushing_tds	'15 Julio Jones / '09 Vincent Jackson	'10 Terrell Owens / '09 Steve Smith	1st	1st	258
5	**Low-Usage Backup Receivers**	rushing_tds / carries	targets / receiving_yards	'09 Brandon Jones / '09 Jerome Simpson	'17 Markus Wheaton / '09 Limas Sweed	5th	3rd	84

Table 5: *Description of Running Back Clusters*

HB Cluster	Description	Best Relevant Stats	Worst Relevant Stats	Best Players in Cluster (highest WPA)	Worst Players in Cluster (lowest WPA)	WPA Rank (mean wpa)	Win Percentage Rank (mean winpct)	Number of Players
1	Low Usage Third Down Receiving Backs	Yards_per_Car / rec_tds	rushing_yards / Carries	'17 Robert Turbin / '11 Jacob Hester	'17 Eddie Lacey / '16 Dwayne Washington	2nd	5th	250
2	High Usage Do-it-all Running Backs	Carries / rushing_yards	Yards_per_Car / TD_per_Car	'10 James Charles / '12 Adrian Peterson	'15 Frank Gore / '11 Cedric Benson	4th	1st	351
3	Low Usage Goal-line Running Backs	TD_per_Car / Yards_per_Car	rushing_tds / rec_tds	'10 Javarris James / '16 Mike Gillislee	'12 Ryan Williams / '13 Andre Brown	3rd	2nd	400
4	Low Usage Short-Yardage Running Backs	Yards_per_Car / rec_tds	rec_yards / Targets	'11 Evan Royster / '11 Chris Ivory	'15 Andre Williams / '16 Adrian Peterson	1st	3rd	159
5	Efficient All-Purpose Running Backs	Targets / carries	Yards_per_Car / TD_per_Car	'10 Maurice Jones-Drew / '14 Arian Foster	'15 Antonio Andrews / '10 Jahvid Best	5th	4th	48

Table 6: *Description of Quarterback Clusters*

QB Cluster	Description	Best Relevant Stats	Worst Relevant Stats	Best Players in Cluster (highest WPA)	Worst Players in Cluster (lowest WPA)	WPA Rank (mean wpa)	Win Percentage Rank (mean winpct)	Number of Players
1	Substandard Balanced Back-Ups	rushing_tds / rushing_yards	passing_yards / attempts	'15 Tony Romo / '11 Matt Flynn	'16 Jared Goff / '12 Ryan Lindley	4th	4th	178
2	Well-Balanced Back-Ups	rushing_tds / rushing_yards	interceptions / attempts	'16 Jacoby Brissett / '14 Connor Shaw	'10 Rusty Smith / '10 David Carr	5th	3rd	93
3	Agile Pocket Passer Back-Ups	rushing_tds / rushing_yards	attempts / passing_yards	'16 Tony Romo / '12 Shaun Hill	'14 Matt Schaub / '13 Tyrod Taylor	3rd	2nd	23
4	Game Managers	interceptions / attempts	rushing_tds / rushing_yds	'15 Andy Dalton / '11 Matt Schaub	'09 Mark Sanchez / '09 Kerry Collins	2nd	5th	133
5	Productive Super Star Quarterbacks	attempts / passing_yards	rushing_tds / rushing_yds	'11 Tom Brady / '13 Manning	'14 Blake Bortles / '11 Blaine Gabbert	1st	1st	218

Results and Conclusions:

When looking at the tables above, there are a couple of conclusions that can come from this. Productive Super Star Quarterbacks are clearly the best when wanting your team to perform well. These Quarterbacks are above average in everything, such as, passing yards, rushing yards, rushing touchdowns, etc. The group ranges from Quarterbacks like Tom Brady and Peyton Manning to Quarterbacks like Michael Vick and Cam Newton.

For the receivers there are two groups that stand out in terms of average WPA. Those two groups are the High-Usage Red-Zone Threats and Middle of the Field Threats. The middle of the field threats are above average in receiving yards and targets but were below average in receiving touchdowns. This means that this cluster group has guys that can get open when needed to since they get a lot of targets as well as receiving yards. However, the group of receivers that gives a team their best chance to win are the High-Usage Red-Zone Threats. These receivers are above average in receiving yards, receiving touchdowns, targets, and even yards after catch. This cluster group really is just efficient at scoring touchdowns and helping their team win. This cluster group contained guys such as Julio Jones, Vincent Jackson, Steve Smith, and Terrell Owens.

When looking at the clusters from the Running Back data there doesn't seem to be a clear cluster that gives a team a better chance at winning. In terms of average WPA the best cluster group were the Low Usage Short-Yardage Running Backs. However, these players were below average in every statistic that was observed. Since this is the case a couple of conclusions could be made from this. One is that the trend of multiple running back system in the NFL could be a successful one. According to my results this is the case, since the cluster with the highest mean WPA contained backs that didn't get many touches, but when they did it resulted in increasing a teams win probability. This means that GM's and Coaches should try and focus on spending their money elsewhere, rather than the Running Back position. This is because, based on my results, all you need is a guy that can punch it in the end zone, or someone that can get that extra one yard for the first down.

There are a number of improvements I could have done to improve this research. One is that the data could've been filtered so that the only players observed were players with a minimum number of targets, attempts, and carries, in order for the players in the data to be just starters. This would've made it easier to decipher and define the clusters more clearly.

Another way this analysis could have been improved was to include other player stats. Some stats I would've liked to have were height, weight, a speed rating, and age. This would have helped define some of the clusters more clearly and would have given a more in-depth analysis in terms of defining clusters that perform well.

References:

1. 2009-2017 NFL Standings & Team Stats. (n.d.). Retrieved from https://www.pro-football-reference.com/years/2017/index.htm
2. Alboukadel, Ahmed, J., & Muzafferiyet, E. (2019, December 25). 5 Amazing Types of Clustering Methods You Should Know. Retrieved from https://www.datanovia.com/en/blog/types-of-clustering-methods-overview-and-quick-start-r-code/
3. Anders Drachen Anders Drachen, & Drachen, A. (2020, November 05). Introducing Clustering I: Behavioral Profiling for Game Analytics. Retrieved from https://gameanalytics.com/blog/introducing-clustering-behavioral-profiling-gameanalytics.html
4. Atkinson, B. (2019, August 12). Visualizing Different NFL Player Styles. Retrieved from https://towardsdatascience.com/visualizing-different-nfl-player-styles-88ef31420539
5. Barnwell, B. (2017, July 21). The NFL stats that matter most. Retrieved from https://www.espn.com/nfl/story/_/id/20114211/the-nfl-stats-matter-most-2017-offseason-bill-barnwell
6. Barra, A. (2018, November 16). The Most Important Stat in Football. Retrieved from https://www.stadiumtalk.com/s/most-important-stat-in-football-47ea57ca1f5249d9
7. Benoit, A. (2019, March 20). The most important trait at every position in football. Retrieved from https://www.si.com/nfl/2019/03/20/nfl-draft-evaluation-most-important-trait
8. Bosch, J. & Kalman, S. (n.d.). NBA Lineup Analysis on Clustered Player Tendencies: A new approach to the positions of basketball & modeling lineup efficiency. Retrieved from https://www.sloansportsconference.com/research-papers/nba-lineup-analysis-on-

clustered-player-tendencies-a-new-approach-to-the-positions-of-basketball-modeling-lineup-efficiency

9. Brecque, C. (2021, April 02). Clustering ga Players. Retrieved from https://towardsdatascience.com/clustering-fifa-players-4408a0954cb4
10. Burke, N. (2019, December 05). Clustering Basketball Players by Position. Retrieved from https://rpubs.com/nburke2/636812
11. C., B. (2014, January 24). College football's 5 most important stats. Retrieved from https://www.footballstudyhall.com/2014/1/24/5337968/college-football-five-factors
12. Chahrouri, E. E. (2018, July 03). QB Clustering – PFF Forecast's way of classifying quarterbacks: NFL News, Rankings and Statistics. Retrieved from https://www.pff.com/news/pro-forecast-pff-qb-clustering-better-way-to-evaluate-qb-performance
13. DaSilva, C. (2017, May 31). Ranking the importance of every NFL position, from QB to long snapper. Retrieved from https://www.foxsports.com/nfl/gallery/nfl-position-importance-ranking-value-every-player-053117
14. Eniyei, Hi, K. O., Kassambara, Ranjit Singh 05 Oct 2019 https://cran.r-project.org/web/packages/factoextra/factoextra.pdf it is written below 0.5 is highly clusterable but in this it is 0.8 how it is highly clusterable Reply, Singh, R., P, I., . . . Yan, H. (2018, October 21). Assessing Clustering Tendency. Retrieved from https://www.datanovia.com/en/lessons/assessing-clustering-tendency/
15. Friscojosh. (2020, January 09). What The NFL's New Pass-Defense Metric Can - And Can't - Tell Us. Retrieved from https://fivethirtyeight.com/features/what-the-nfls-new-pass-defense-metric-can-and-cant-tell-us/

16. Godfrey, K., Kassambara 08 Dec 2018 With the elbow method (method = “wss”), Kassambara, Romero, J., Lucía, Fit, P., . . . Roth, G. (2018, October 21). Determining The Optimal Number Of Clusters: 3 Must Know Methods. Retrieved from https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/
17. Goldberg, S. (2020, March 03). Comparing Players: Clustering and Style of Play. Retrieved from https://www.americansocceranalysis.com/home/2020/3/3/clustering
18. Green E., Menz M., Benz L., Zanuttini-Frank, G., & Bogaty M. (2016, April). Clustering NBA Players.
Retrieved from https://sports.sites.yale.edu/clustering-nba-players
19. Hussain, H. (2019, November 08). Using K-Means Clustering Algorithm to Redefine NBA Positions and Explore Roster Construction. Retrieved from https://towardsdatascience.com/using-k-means-clustering-algorithm-to-redefine-nba-positions-and-explore-roster-construction-8cd0f9a96dbb
20. Justin. (2015, September 29). NBA Positions by Clustering. Retrieved from https://fansided.com/2015/09/29/nba-positions-by-clustering/
21. Kacsmar, S. (2017, October 03). 6 NFL Stats Fans and Media Don't Understand. Retrieved from https://bleacherreport.com/articles/1309665-6-nfl-stats-casual-fans-media-just-dont-understand
22. Keim, J. (2015, May 22). Which defensive stats matter most in helping teams win? Retrieved from https://www.espn.com/blog/nflnation/post/_/id/169435/which-defensive-stats-matter-most-in-helping-teams-win

23. Kilitcioglu, D. (2018, October 26). Hierarchical Clustering and its Applications. Retrieved from https://towardsdatascience.com/hierarchical-clustering-and-its-applications-41c1ad4441a6#:~:text=Conclusion,far%20apart%20data%20points%20are.
24. Lee, J. (2020, June 18). Grouping Soccer Players with Similar Skillsets in FIFA 20: Part 1: K-Means Clustering. Retrieved from https://towardsdatascience.com/grouping-soccer-players-with-similar-skillsets-in-fifa-20-part-1-k-means-clustering-c4a845db78bc
25. Leisy, R. (2020, July 05). Who is the next Julio Jones? Retrieved from https://medium.com/@leisyridley/who-is-the-next-julio-jones-b7de48aebe18
26. Lesmeister, C. (2014, December 30). Cluster Analysis of the NFL's Top Wide Receivers: R-bloggers. Retrieved from https://www.r-bloggers.com/cluster-analysis-of-the-nfls-top-wide-receivers/
27. Morse, S. (n.d.). Clustering NFL Wide Receivers by Individual Play Distributions. Retrieved from https://stmorse.github.io/journal/clustering-nfl-players-by-play-distributions.html
28. Neill, B. (2020, October 29). Has receiver value overtaken that of a quarterback?: PFN. Retrieved from https://www.profootballnetwork.com/value-by-position-have-receivers-replaced-quarterbacks-as-most-valuable/
29. Pyne, S. (2017, April 24). Quantifying the Trenches: Machine Learning Applied to NFL Offensive Lineman Valuation. Retrieved from https://core.ac.uk/download/pdf/84114251.pdf
30. Ryurko. (n.d.). Ryurko/nflscrapR-data. Retrieved from https://github.com/ryurko/nflscrapR-data

31. Seif, G. (2021, January 25). The 5 Clustering Algorithms Data Scientists Need to Know. Retrieved from https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68
32. Shalizi, C. (2009, September 14). Distances between Clustering, Hierarchical Clustering. Retrieved from https://www.stat.cmu.edu/~cshalizi/350/lectures/08/lecture-08.pdf
33. Shook, N. (2020, July 22). Ten best-performing NFL O-lines by expected rushing yards. Retrieved from https://www.nfl.com/news/ten-best-performing-nfl-o-lines-by-expected-rushing-yards

Appendix:

Figure 4: *Elbow Method for Wide Receiver Dataset*

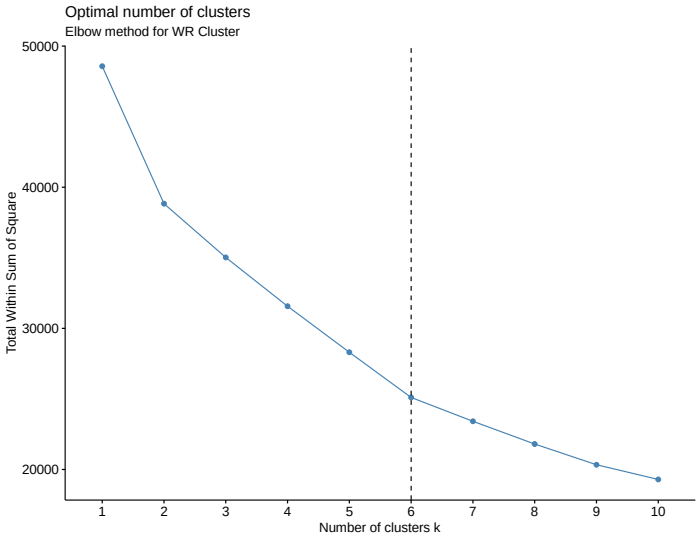


Figure 5: *Silhouette Method for Wide Receiver Dataset*

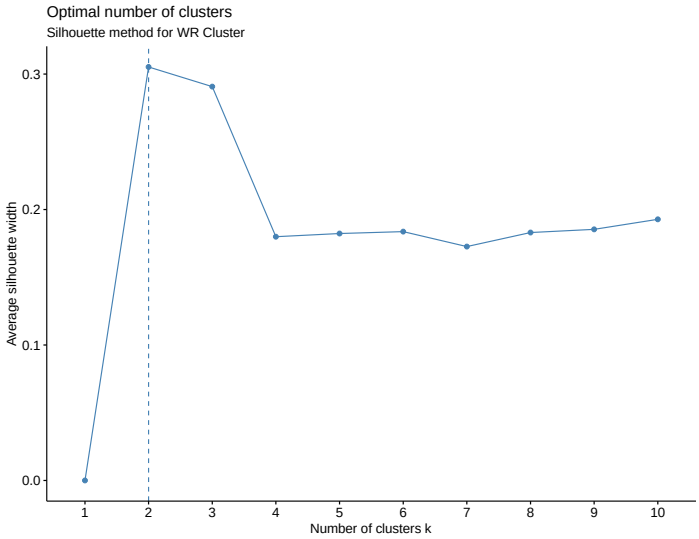


Figure 6: *Gap Statistic Method for Wide Receiver Dataset*

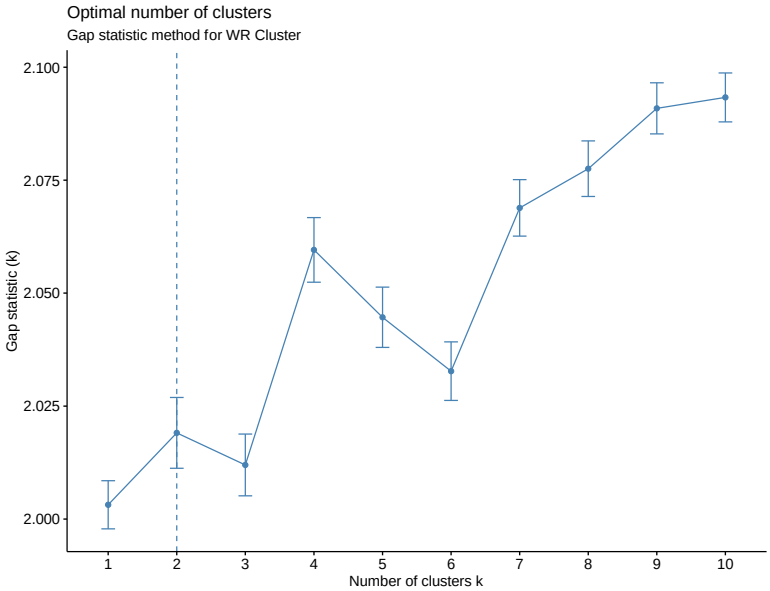


Figure 7: *Elbow Method for Running Back Dataset*

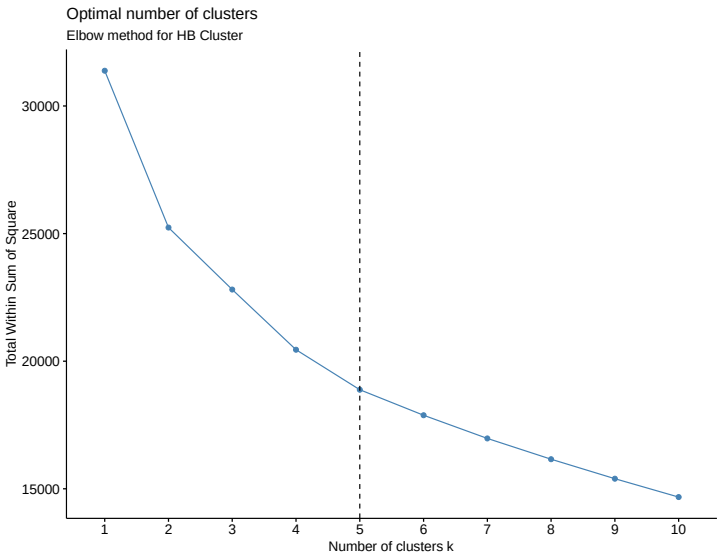


Figure 8: *Silhouette Method for Running Back Dataset*

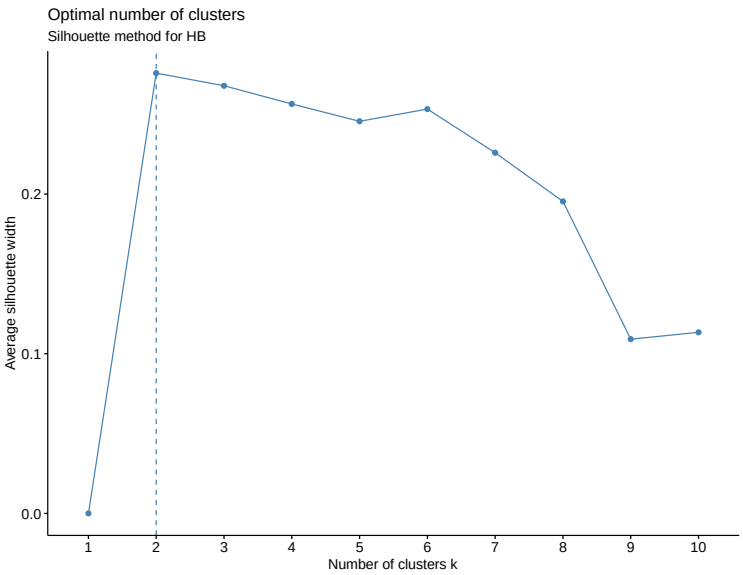


Figure 9: *Gap Statistic Method for Running Back Dataset*

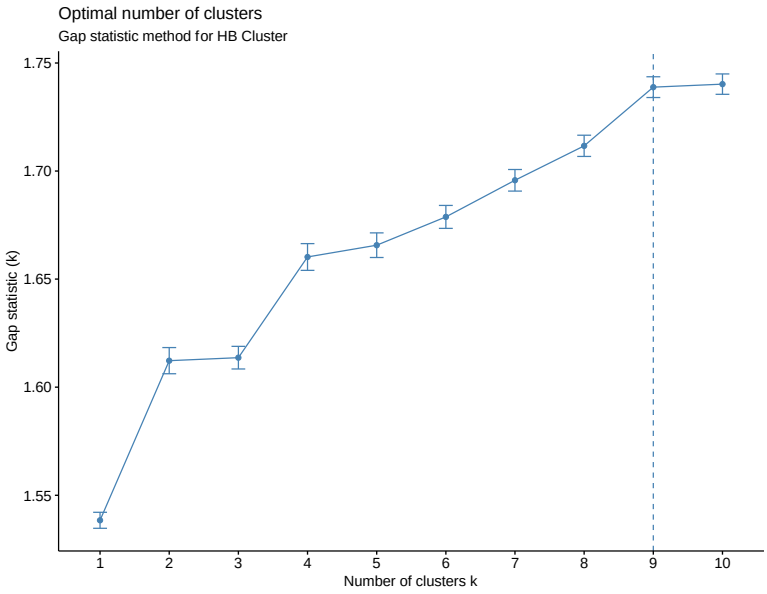


Figure 10: *Elbow Method for Quarterback Dataset*

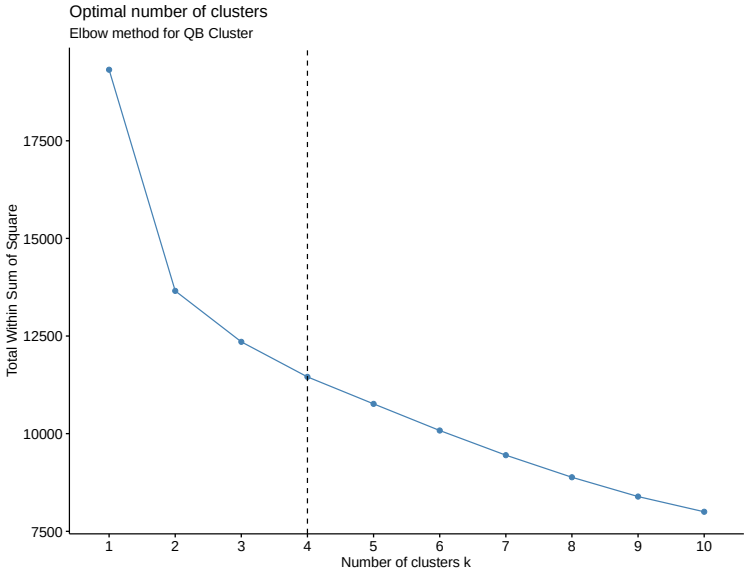


Figure 11: *Silhouette Method for Quarterback Dataset*

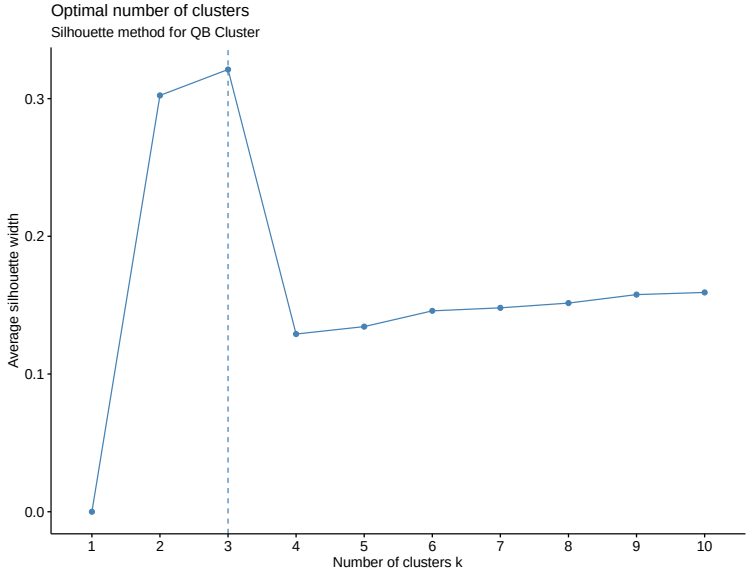


Figure 12: *Gap Statistic Method for Quarterback Dataset*

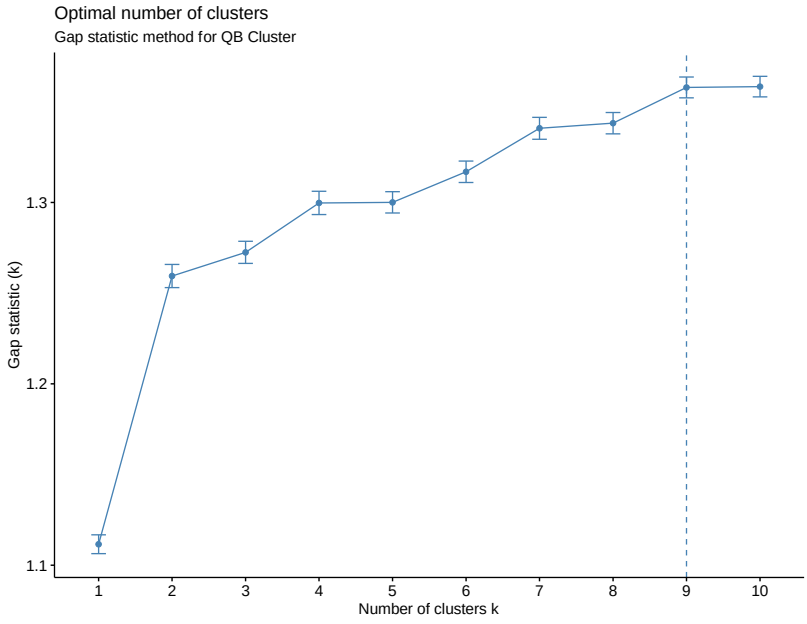


Figure 22: *Breakdown of Wide Receiver Cluster 1*

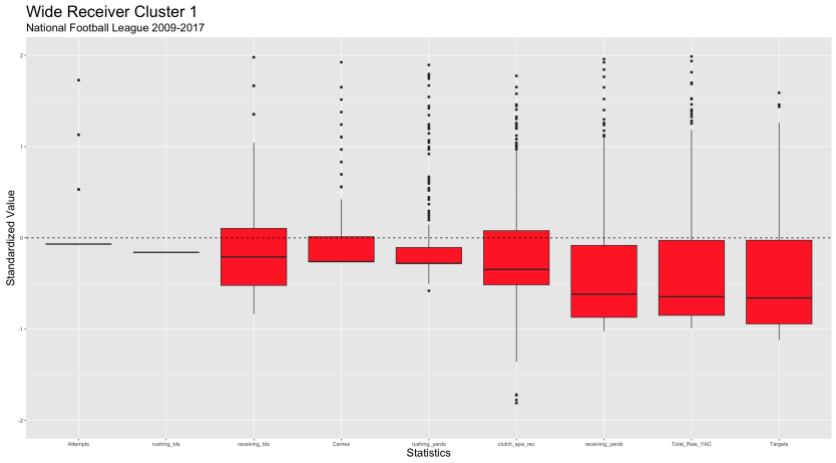


Figure 23: *Breakdown of Wide Receiver Cluster 2*

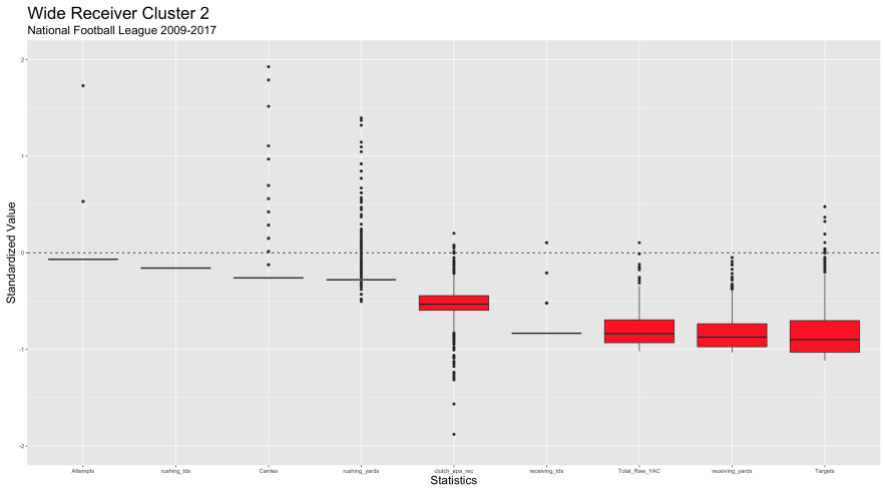


Figure 24: *Breakdown of Wide Receiver Cluster 3*

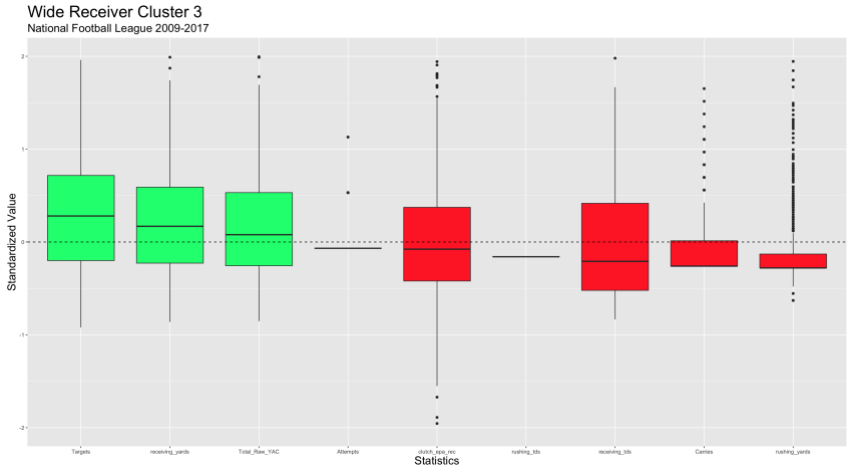


Figure 25: *Breakdown of Wide Receiver Cluster 5*

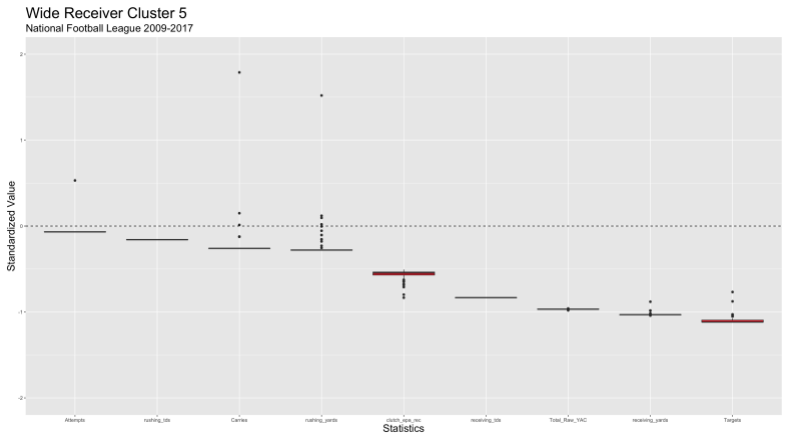


Figure 26: *Breakdown of Running Back Cluster 1*

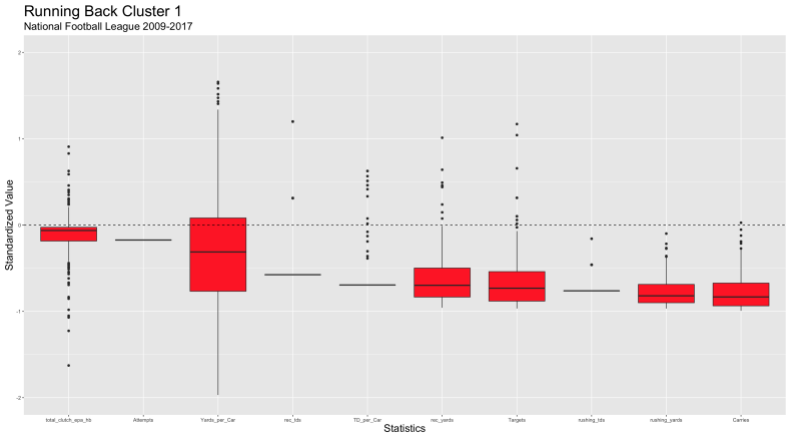


Figure 27: *Breakdown of Running Back Cluster 2*

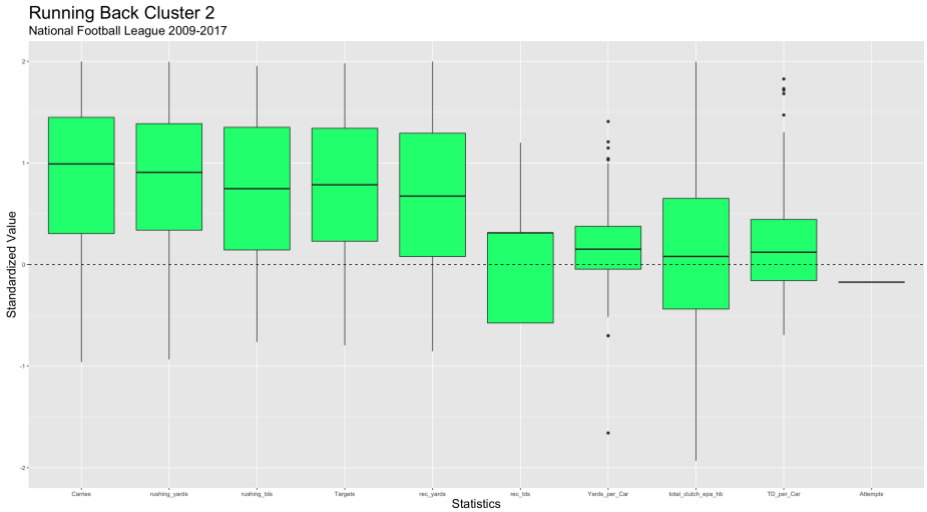


Figure 28: *Breakdown of Running Back Cluster 3*

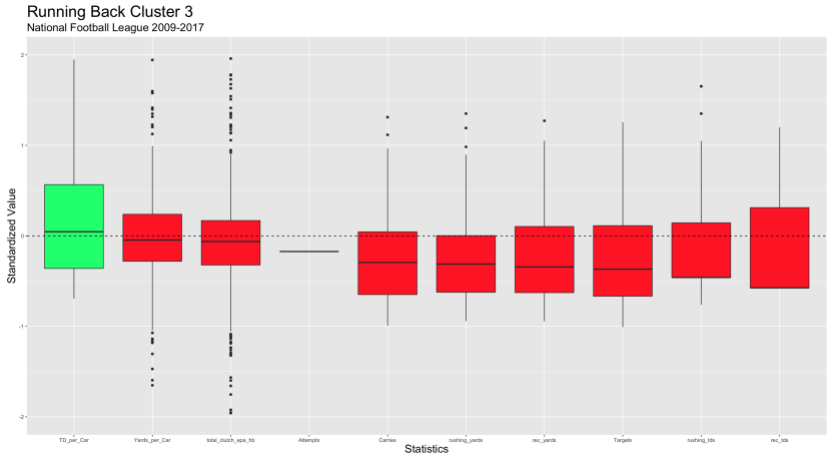


Figure 29: *Breakdown of Running Back Cluster 5*

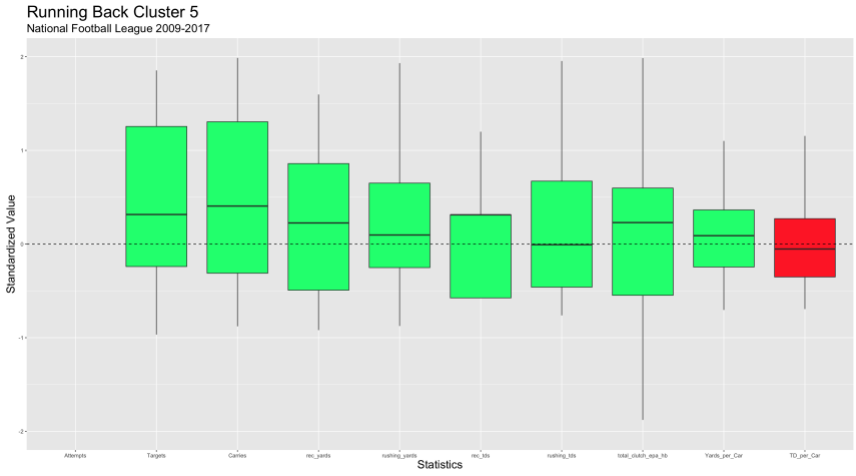


Figure 30: *Breakdown of Quarterback Cluster 1*

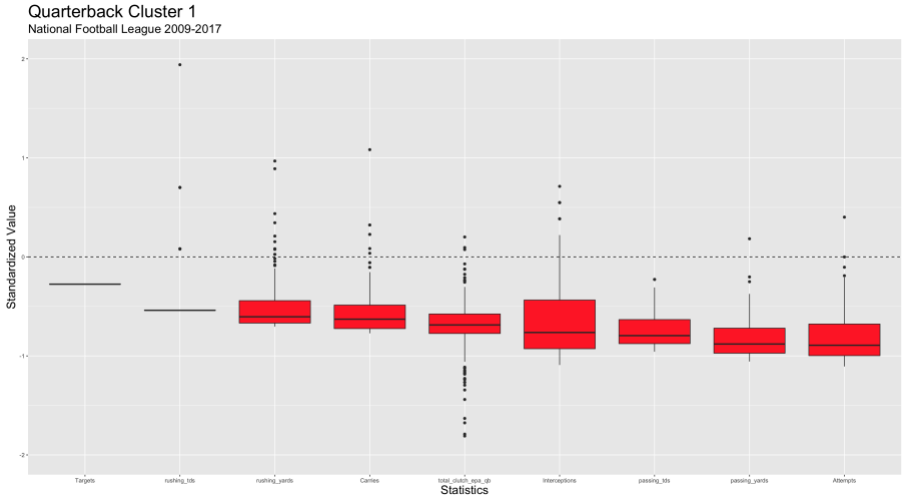


Figure 31: *Breakdown of Quarterback Cluster 2*

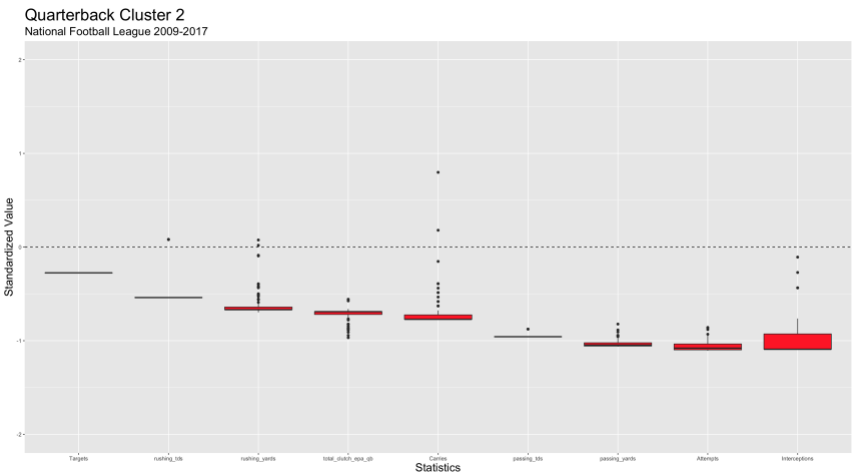


Figure 32: *Breakdown of Quarterback Cluster 3*

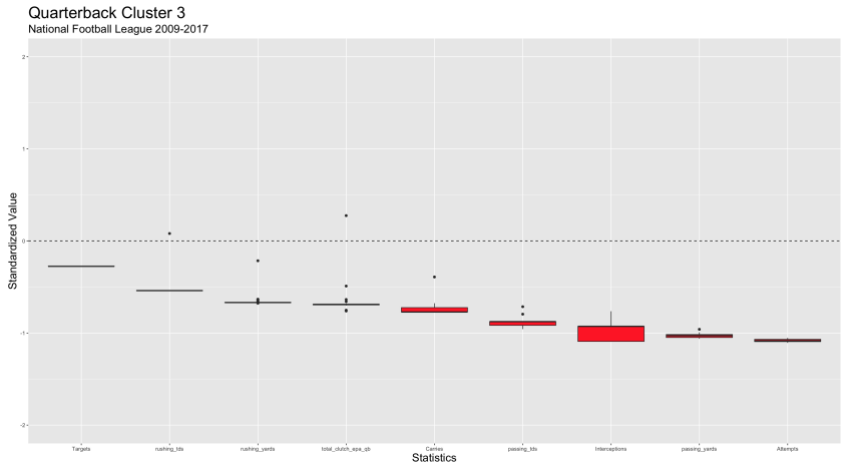


Figure 33: *Breakdown of Quarterback Cluster 4*