Canton, Ohio Forecasting: Using Statistical Analysis and Visualization to Predict Future NFL Hall of Fame Inductees

Matt Sherrick

Abstract The purpose of this report, as the name implies, is to make predictions on what current or recently retired NFL players will end up hearing their names called in Canton, Ohio to be inducted into the Pro Football Hall of Fame. I achieve this by using a variety of visualizations to explore patterns between current and Hall of Fame players, as well as using different methods of clustering to group the players together. As an avid sports fan, I also utilize my personal sports knowledge and outside data, like player rewards, injury history, and Super Bowl victories, to help make my predictions more accurate. Using the results from the clustering methods, as well as considering external factors that were not included in the dataset, I make four to five predictions for quarterbacks, running backs, and wide receivers that will end up in the Hall of Fame following the conclusion of their careers.

Contents

Section 1 : Motivation

For my thesis, I am analyzing a large dataset of NFL statistics that includes career stats for all position players as well as game logs for each player. I have had a strong passion for sports all my life. My interest stemmed from all the times I've seen random stats show up on ESPN that made me ask myself, "How do they come up with that?" That is how I knew I wanted to do something sports related; I wanted to combine my love of sports with my knowledge of statistics for my thesis and see what insights I can find working with a dataset from one of the four major professional sports leagues. After exploring various datasets on kaggle.com, I chose to analyze a dataset titled NFL Statistics [8], as it included all the variables that would help me answer my research question. My goal is to observe the career statistics of Hall of Fame players and compare them to current or recently retired players to be able to predict who will end up in the Hall of Fame based on multiple criteria. For my report, I specifically look at quarterbacks, running backs, and wide receivers. Some of the variables that I look at that indicates success include total yards and touchdowns throughout one's career. To be able to decide who among the current players in the dataset will end up in the Hall of Fame, I can look at patterns for each statistic to determine if those players are on pace to tie or pass the career numbers set by Hall of Famers. I can also use statistical methods like clustering to see if the players of interest should be placed in the same category as the current Hall of Fame players.

Section 2 : Data

The dataset [8] itself includes three different types of smaller datasets: basic stats, career stats, and game logs. The basic statistics file primarily includes information about each player in the dataset such as their number, position, current team, age, and more. The second group, career stats, includes all the statistics separated by category for each player. This includes career numbers in passing, rushing, receiving, tackling, field goals, and more. While each player has their own primary position, there are some drawn up plays where a player can record stats in different categories by lining up as a position other than their own. For instance, in the passing stats dataset, some wide receivers are included because there are often trick plays that involved the receiver throwing the ball rather than the quarterback. There are also some players who are labeled as defensive position players that appear in the passing, receiving, and rushing datasets. For the purpose of my report, I will be working with the Career Stats Passing, Career Stats Rushing, and Career Stats Receiving files to predict future Hall of Fame inductees who played quarterback, running back, or wide receiver. It is also important to note that this dataset was created and last updated in 2017, so the statistics for every player is recorded up to the conclusion of the 2016 season.

2.1 : Variables

Quarterbacks

To start off, I looked at the Career Stats Passing dataset, which includes 8,525 observations and 23 columns of the yearly stats of every player who attempted at least one pass in that season. The other variables include their team, position, passes completed, passes attempted, completion percentage, passing yards, touchdowns thrown, interceptions, and passer rating. The majority of the variables are categorized as character variables, so to work directly with the statistics I will convert the variables of interest to numeric ones.

Running Backs

In the Career Stats Rushing dataset, there are 17,507 entries and 16 columns that list every season a player has rushed at least once. Along with running backs, these data entries can include quarterbacks, receivers, and even punters or kickers. Similarly to the Passing Stats dataset, the entries include the player's name, team, and year. The other variables include rushing attempts, rushing yards, yards per carry, and fumbles.

Wide Receivers

In the Career Stats Receiving dataset, we have 18,128 entries and 19 with a similar layout to the other two datasets in that it lists every player that caught at least one pass throughout the season. In addition to the

name and team, this dataset includes the variables receptions, receiving yards, yards per reception, receiving touchdowns, and fumbles.

2.2 : Cleaning and Processing

Some of the issues with all the datasets were that not every player in the league's history are included. When working with the career stats in passing dataset, I researched all the current quarterbacks in the Hall of Fame. Then, when looking at the dataset, only a select amount of Hall of Fame inductees were included in the dataset. The same issue arose looking for all the current Hall of Famers in the rushing and receiving stats datasets. So for my research, I only considered Hall of Fame players that were included in the datasets.

Another issue was that a lot of the entries in the dataset had NA under their position. This led to uncertainty whether or not, for instance, all the NA entries in the passing dataset were all quarterbacks since other position players can throw passes too. The same problem existed in the rushing and receiving datasets For that reason, I excluded all those entries so I could specifically focus on players that were labeled as quarterbacks, running backs, and wide receivers.

2.2.1 : Quarterbacks

Starting with the quarterbacks, our first step is to take all the missing entries that are indicated with "–" and replace it with NA. That way, we can use the `na.omit` command [14] to effectively remove all the missing values in our dataset. As indicated in the previous section, I only want to include the quarterbacks that are labeled "QB" under position to avoid confusion and the potential of including non-quarterback players. Next, I realized many of the variables were not of the class numeric when importing the dataset. Using the `as.numeric` command [14], we convert all the variables of interest to numeric variables. For quarterbacks, I want to look at passing yards, touchdowns, pass attempts and completions, games played, and interceptions.

2.2.2 : Running Backs

We follow the same procedure for the Career Stats Rushing dataset creating NA values in the columns of interest and changing them to be numeric variables. The variables I will be considering are games played, rushing yards, rushing attempts, yards per carry, touchdowns, and fumbles.

2.2.3 : Wide Receivers

Similarly, the same steps are taken for the receiving dataset. For this, our variables of interest are games played, receptions, receiving yards, yards per reception, touchdowns, and fumbles.

2.3 : Preparing the Data

To prepare our data for later methods, we have to create subsets for each of our 3 datasets based on the criteria we want.

2.3.1: Quarterbacks

Starting with the passing data, in order to create a dataset I can use to compare the stats of active and Hall of Fame quarterbacks, I use the `aggregate` [14] command to sum up the passing yards, touchdowns, pass attempts and completions, games played, and interceptions and link these columns to the name of each player. I added a categorical variable called Hall of Fame, which held the value 0 or 1, with 1 indicating that the player is inducted into the Hall of Fame.

Initially, we have several entries that have only a handful of pass attempts over the whole season where most of them have NA under their position. Additionally, under games played, these entries had 10 or less games. This means that these rows most likely indicate other position players throwing on trick plays or quarterbacks that did not start. For my research, I want to look at starting quarterbacks only since they have accrued the most stats over the years. To achieve this, I create a subset of the dataset with the two conditions: the position is labeled quarterback, and the total games played in the season exceeds 10.

Next, I searched through the dataset to find all the Hall of Fame players that were included in the dataset and gave them a value of 1 under the Hall of Fame column. I then created another subset with this condition for all the Hall of Fame quarterbacks. With these datasets ready, I combined the two to make an All Quarterbacks dataset with current and Hall of Fame players that I will use for comparison.

2.3.2 : Running backs

I will follow similar steps as what I did with the passing stats dataset. There are over 17,000 entries in the dataset, so in order to prevent the possibility of accidentally including positions other than running backs, we use the `subset` [14] command to create a group of players that are labeled "RB." I will also be subsetting the data into players with more than 100 career rushing attempts so we have a more accurate pool of active players to use. Now we can use the `aggregate` command [14] to sum up the variables of interest, combine the columns, and obtain the dataset we want.

After completing our dataset of active running backs, we now create the dataset of Hall of Fame running backs that were included in the dataset.

Our final step to prepare the data is to now add the Hall of Fame indicator and combine our two datasets.

2.3.3 : Wide Receivers

The same procedure will be followed for our wide receivers. Using the `subset` [14] command again, I subset the data by following the two conditions: the position is "WR" and over 50 receptions. That way, we can make sure we only include wide receivers in our dataset and if they have over 50 catches, we'll know they have been playing for a few years. This will also take out a lot of the entries that only had 10 or fewer receptions. After that, we again go back to the `aggregate` command [14] to sum up our variables of interest

Now we can do the same with the Hall of Fame receivers included in our dataset.

Finally, like the last two datasets, we add a Hall of Fame indicator to the active and Hall of Fame receivers then combine the two.

We now have a completed dataset for the quarterbacks, running backs, and receivers. Our finalized quarterback dataset has 76 entries with 8 variables: Name, Games Played, Passing Yards, Passes Attempted, Passes Completed, Touchdowns, Interceptions, and Hall of Fame. The running back dataset has 86 entires with 8 variables: Name, Games Played, Rushing Attempts, Rushing Yards, Yards Per Carry, Touchdowns, Fumbles, and Hall of Fame. Lastly, the receiver dataset has 102 rows and 8 varibales: Name, Games Played, Receptions, Receiving Yards, Yards Per Reception, Touchdowns, Fumbles, and Hall of Fame. Moving on to the next section, we can do some EDA to explore the relationships between certain variables and analyze the differences in these statistics between the current players and Hall of Famers.

Section 3 : Exploratory Data Analysis

My next step is to explore different visualizations to observe the types of patterns exhibited in the career statistics of Hall of Fame players so that I could make comparisons to active players. My goal is to observe the total numbers in each variable of interest, such as total touchdowns, to see if there are current players on pace to achieve the same statistics that only the most elite positions players have done.

3.1 : Quarterbacks

We begin by looking at our quarterbacks. We can use the `ggplot2` [9] package and analyze a variety of plots to observe the relationships in statistics between the active quarterbacks and the Hall of Fame quarterbacks.

3.1.1 : ggplot

Figure 1 was achieved whilst performing EDA. In the scatterplot, we can see the total career touchdowns for both the Hall of Fame and other quarterbacks. This gives us an understanding of the criteria necessary

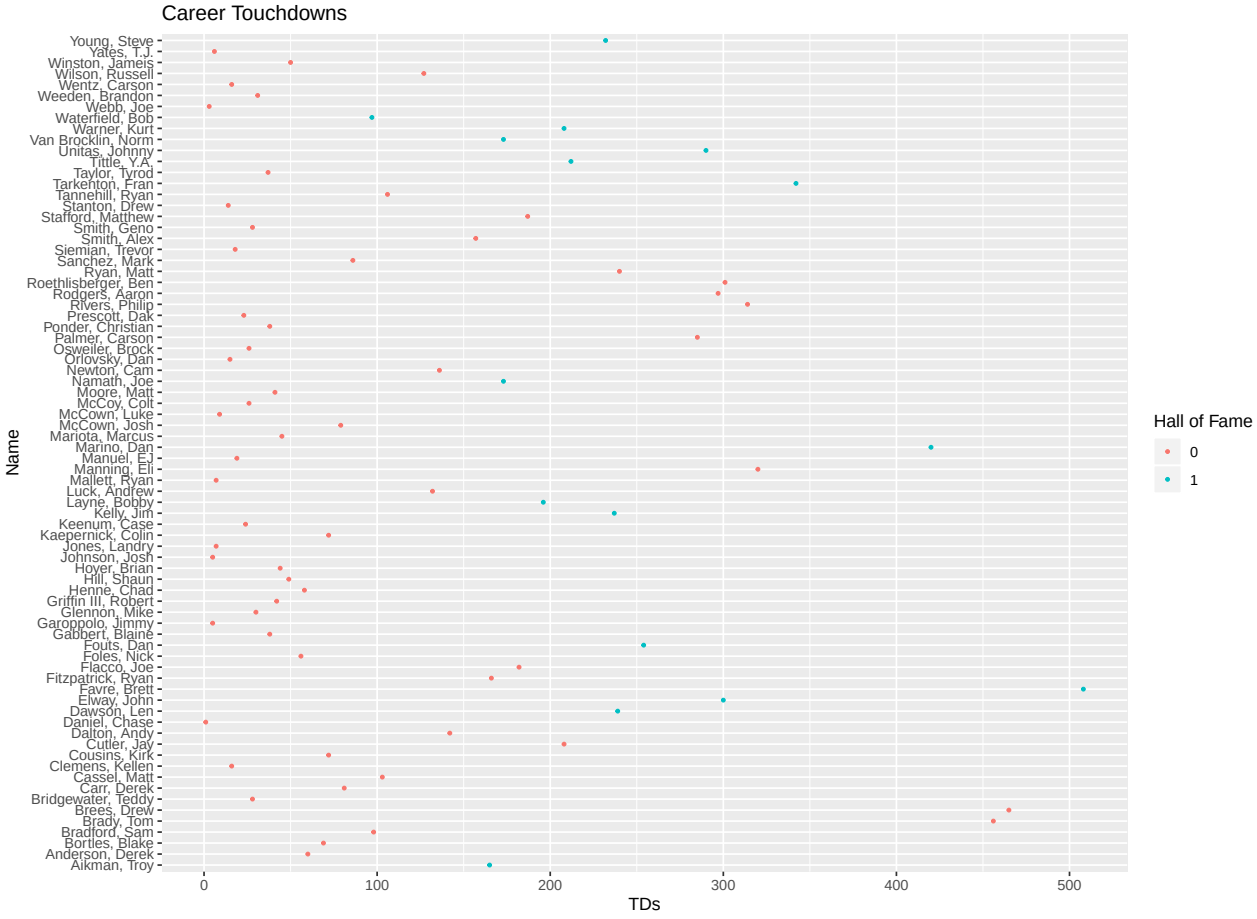

Figure 1: Career Touchdowns for Current and Hall of Fame QBs

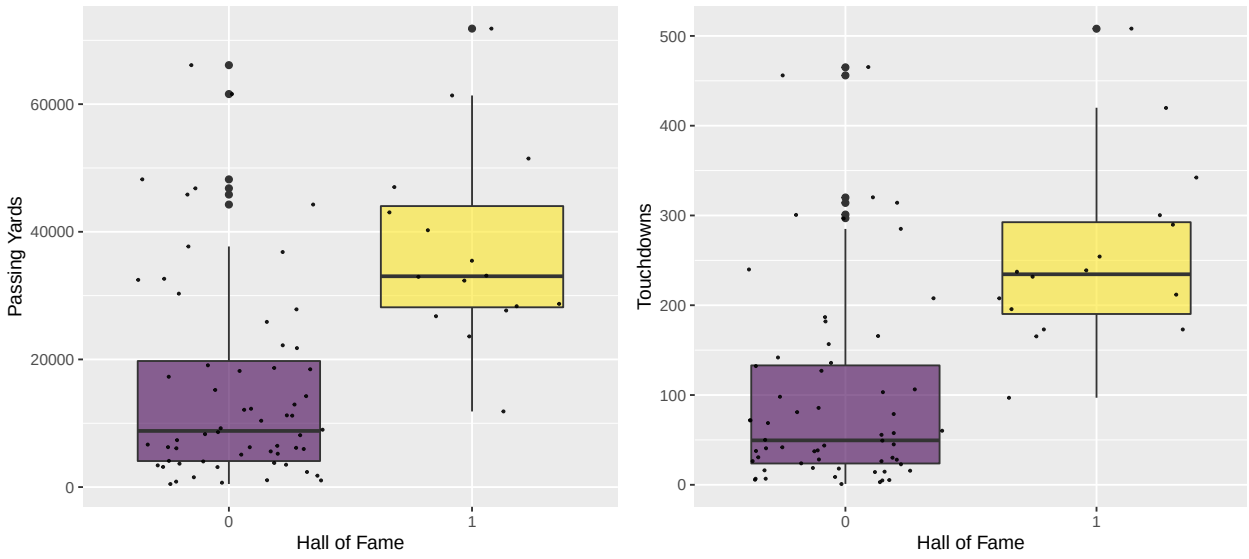

Figure 2: Boxplots of Passing Yards and Touchdowns for active and Hall of Fame quarterbacks

for quarterbacks to be considered for the Hall of Fame. We can see that some current quarterbacks, like Tom Brady and Drew Brees, are already over 400 career touchdowns, which only two current Hall of Fame inductees have achieved. This is a very notable achievement for any quarterback in the league. In sports media today, these two players are arguably the best in their position and have already been considered in Hall of Fame discussion. Some quarterbacks that have reached 300 career touchdowns, like Ben Roethlisberger, Eli Manning, and Aaron Rodgers, are also being considered for the Hall of Fame. Out of these five quarterbacks, all of them have won at least one Super Bowl in their careers. While rings are considered a team achievement rather than an individual one, these quarterbacks are considered the leaders of their offense, and it requires a lot of talent to achieve such a feat. I will discuss the ring factor in a later section. To visualize 1 in a cleaner way, we can use boxplots. We'll also consider passing yards as a potential Hall of Fame indicator. Using the packages `gridExtra` [4] and `gtable` [10], we can arrange our ggplots to be displayed together using the `grid.arrange` [4] command. Additionally, for our the plots here, we use the packages `hrbrthemes` [5] and `viridis` [16] for additional themes on the ggplots.

In Figure 2, we can see that there is a significant difference in total passing yards and touchdowns between the two categories. This gives us a general idea of the criteria that needs to be met for quarterbacks to end up in the Hall of Fame. Some of the outliers in the left boxplot for both touchdowns and passing yards include Tom Brady, Drew Brees, and Eli Manning. As we mentioned before, these quarterbacks are among the few players that have been discussed to be future Hall of Famers in today's media.

3.1.2 : Correlation and Deriving Career-Passer Rating

When looking at the Passer Ratings after each season for the Hall of Fame quarterbacks, they all had seasons with high ratings. These were also statistically their best years. I created a subset of the data that just included the numeric measures, created a correlation matrix, and then used the `corrplot` [17] function to create a correlation plot. Looking at Figure 3 at Passer Rating, we see it is positively correlated with Completion Percentage, Passing Yards Per Attempt, and Percentage of TDs per Attempts and negatively correlated with Interception Rate. These four statistics are how the NFL computes the Passer Rating score, which has a maximum of 158.3. Averaging the Passer Ratings for each season a quarterback has played does not equate to their true career rating. To calculate a quarterback's career rating, I will create a variable that computes the Passer Rating using the variables Passes Attempted, Passes Completed, Passing Yards, Touchdowns, and Interceptions.

To get the career passer rating for each quarterback, we use the following formula [18]:

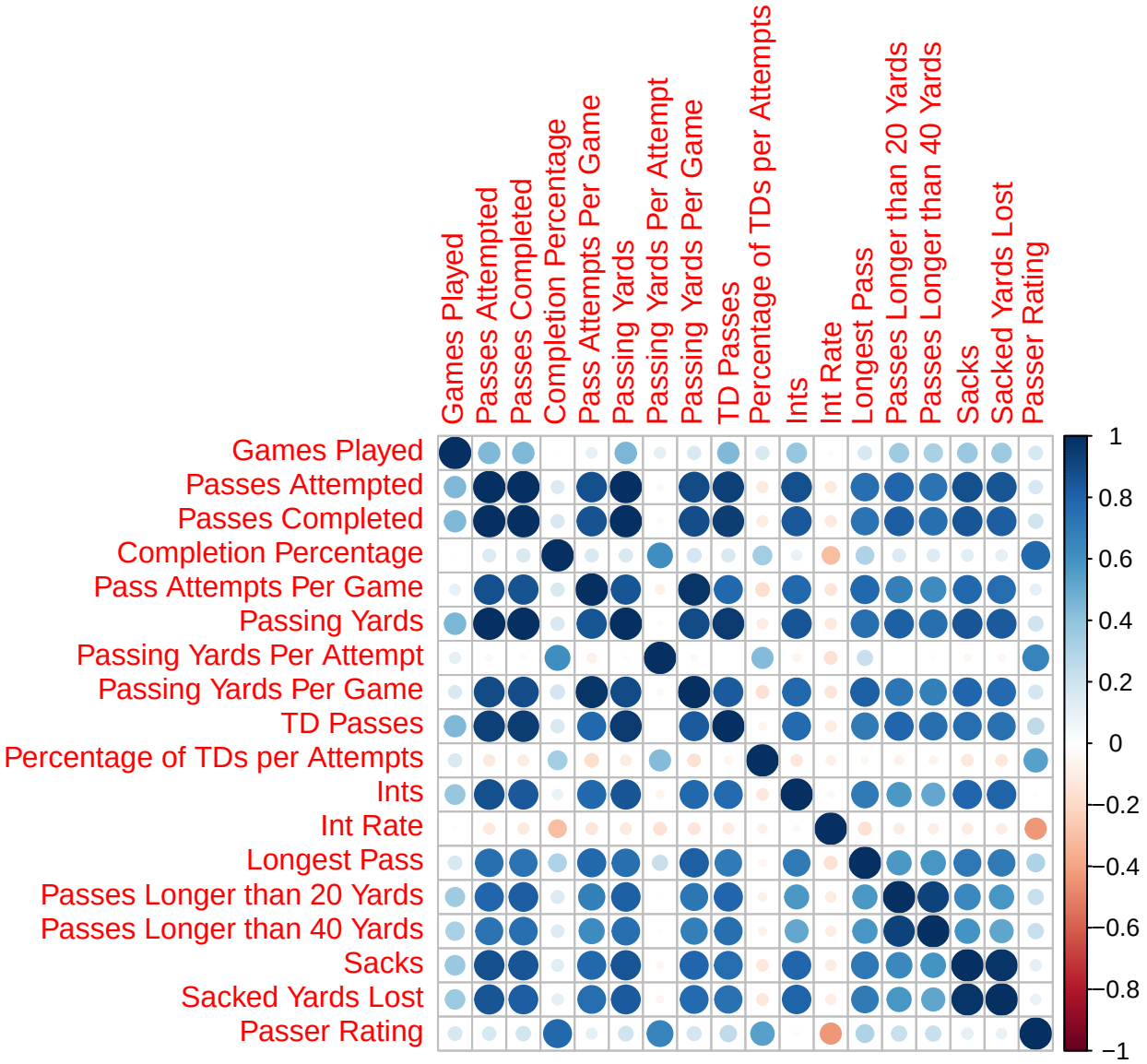

Figure 3: Correlation Plot of Career Passing Stats

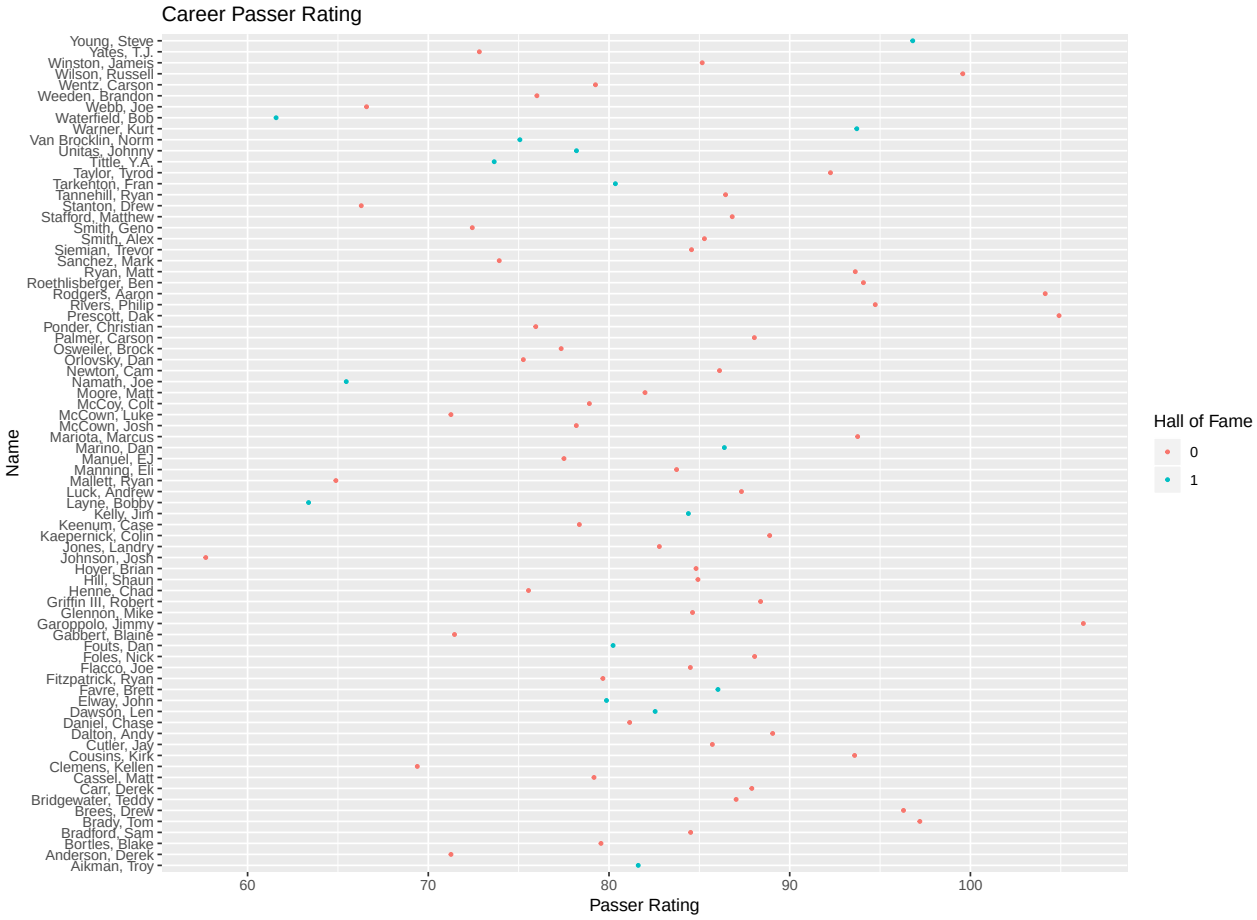

Figure 4: Career Passer Rating for Current and Hall of Fame QBs

$$Career\ Passer\ Rating = 100 \times \left(5 \times \left(\frac{Passes\ Completed}{Passes\ Attempted} - 0.3\right)\right) + 0.25 \times \left(\frac{Passing\ Yards}{Passes\ Attempted} - 3\right) + 20 \times \left(\frac{Touchdowns}{Passes\ Attempted}\right) + 2.375 - \left(25 \times \left(\frac{Interceptions}{PassesAttempted}\right) \div 6\right)$$

With this formula, we add a new column to our dataset, Career Passer Rating. In football today, this statistic is a great indicator of success, so we can look at the passer ratings for all current and Hall of Fame players.

In Figure 4 we see a lot less Hall of Fame players to the right of the graph than current players. There really is no pattern when it comes to Hall of Fame and non Hall of Fame players. This could be due to the fact that passer rating became an official statistic for the NFL in 1973. We can observe this relationship using the `geom_boxplot` [9] command. We again use the packages `hrbrthemes` [5] and `viridis` [16] for additional themes on the ggplots.

From the side by side boxplots in Figure 5, we see that there really is no difference in passer rating between Hall of Fame and non Hall of Fame players as we discovered before. It was worth observing, but we can see that the passer rating statistics really have no significant impact.

3.2 : Running backs

As we saw from our quarterbacks, some of the best predictors of success were total yards and touchdowns. I will also consider yards per carry, as this variable indicates the efficiency of every player as a rusher. Using `ggplot2` [9], I will analyze the same two variables and compare these statistics between the current and Hall

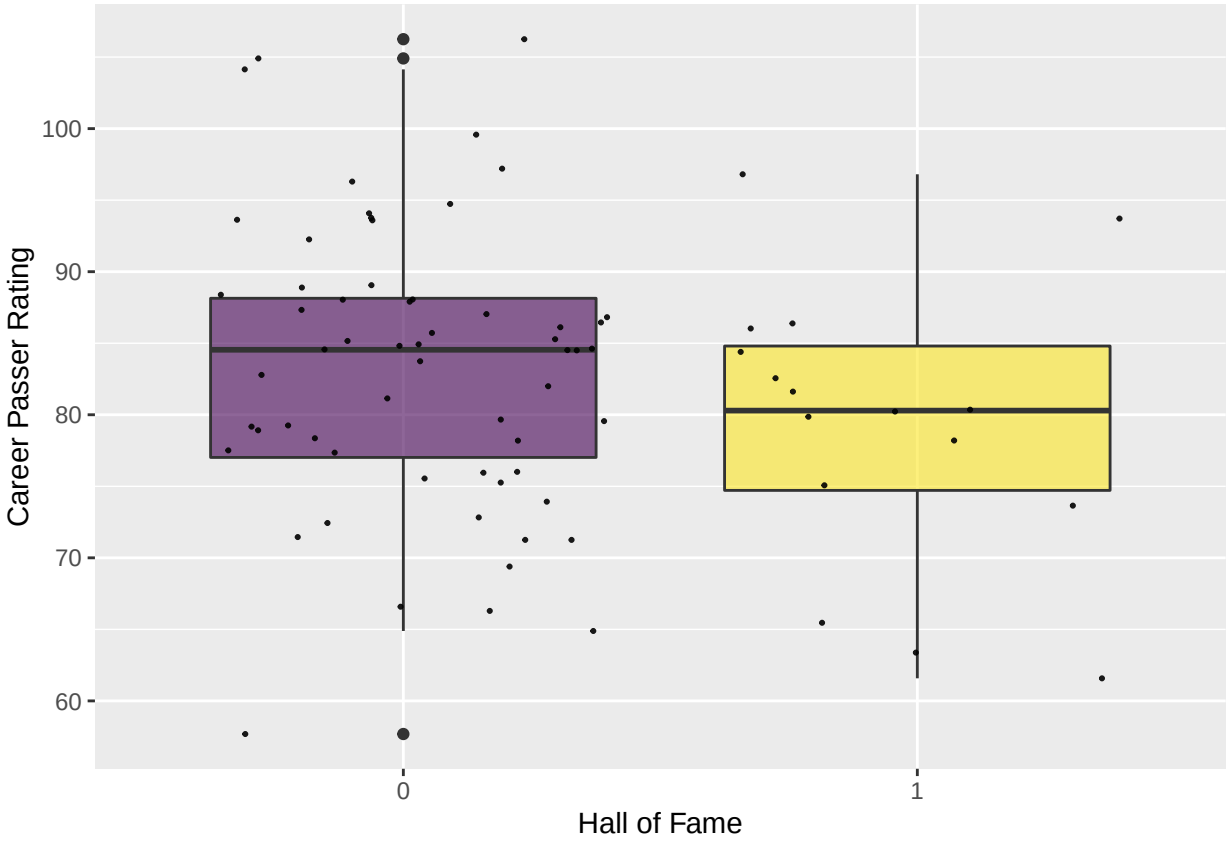

Figure 5: Boxplot of Career Passer Rating for active and Hall of Fame quarterbacks

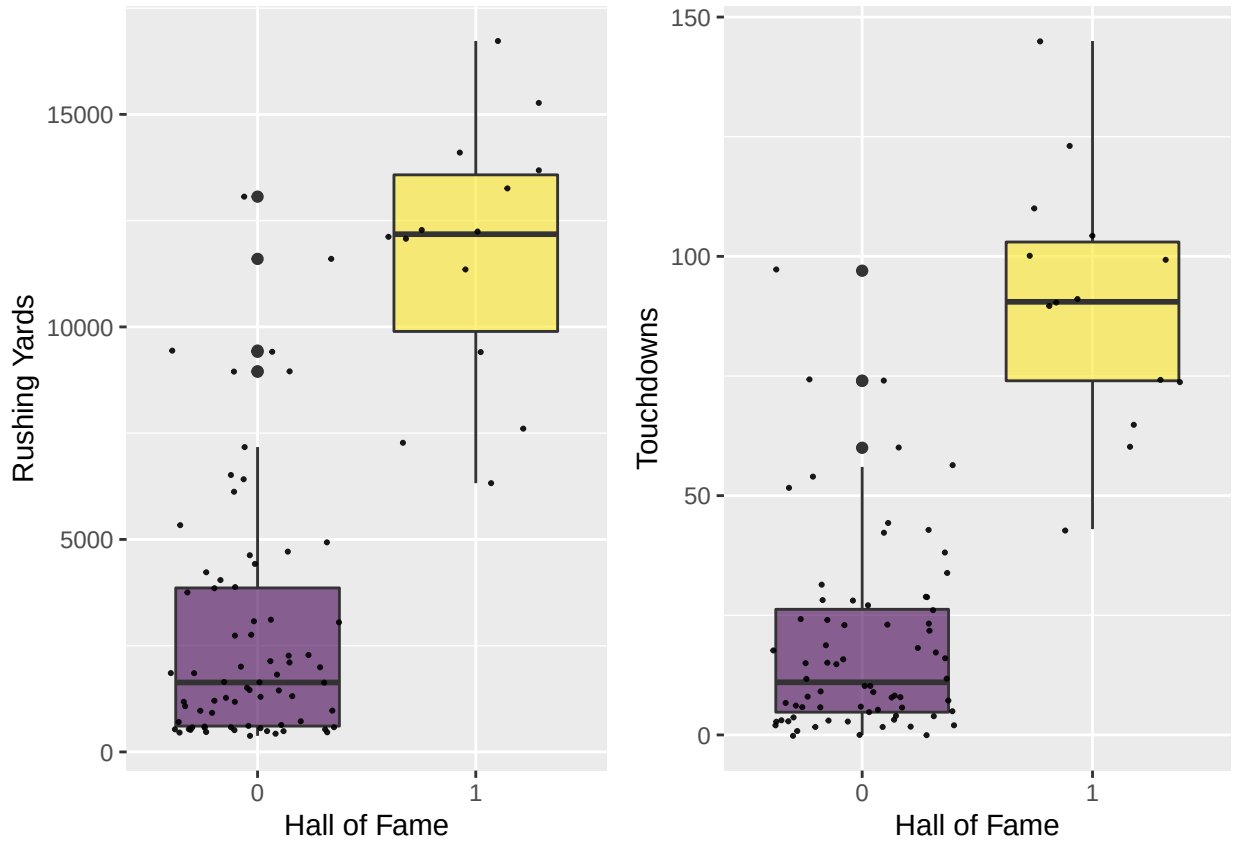

Figure 6: Boxplot of Rushing Yards and Touchdowns for current and Hall of Fame Running Backs

of Fame players. Then, using the `grid.arrange` [4] command again, we can create side-by-side plots.

In Figure 6, we definitely see a big gap in rushing yards and touchdowns for the current and Hall of Fame players. Like before, Hall of Fame players are indicated with a 1 and the current players are indicated by a 0. The outliers for the left boxplot leads us to believe these players are likely on pace to pass or tie the Hall of Fame players' numbers. Looking at our dataset, the two players with over 10,000 rushing yards are Adrian Peterson and Frank Gore. The other two outliers with just under 10,000 yards are Chris Johnson and Matt Forte. As for touchdowns, the three outliers in the left boxplot are Adrian Peterson, Frank Gore, and Marshawn Lynch based on the dataset. Marshawn Lynch recently retired at a young age and shocked the league when he returned for one more season. He is now officially retired for good and is considered one of the most influential power runners in the league.

In Figure 7, there is not really a difference between the Yards Per Carry for the two categories. In fact, most of the current players lay in the same range as the Hall of Fame players with some outliers. Some of the outliers with over 5 yards per carry are players that have only played 15 games or less. This isn't the most ideal sample size, as yards per carry fluctuate a lot throughout a player's career. For this reason, rushing yards and total touchdowns should be the statistics we analyze the most.

3.3 : Wide Receivers

For the wide receivers, I will be considering the same variables of interest as we did with the running backs: total yards, touchdowns, and yards per reception. Similarly, I will be using `ggplot2` [9], along with `hrbrthemes` [5] and `viridis` [16] to produce the visualizations for this section.

In Figure 8, we see the same boxplot pattern as we have seen in the other two positions. We can observe the general criteria set by the Hall of Fame receivers in terms of total yards and touchdowns, again indicated by

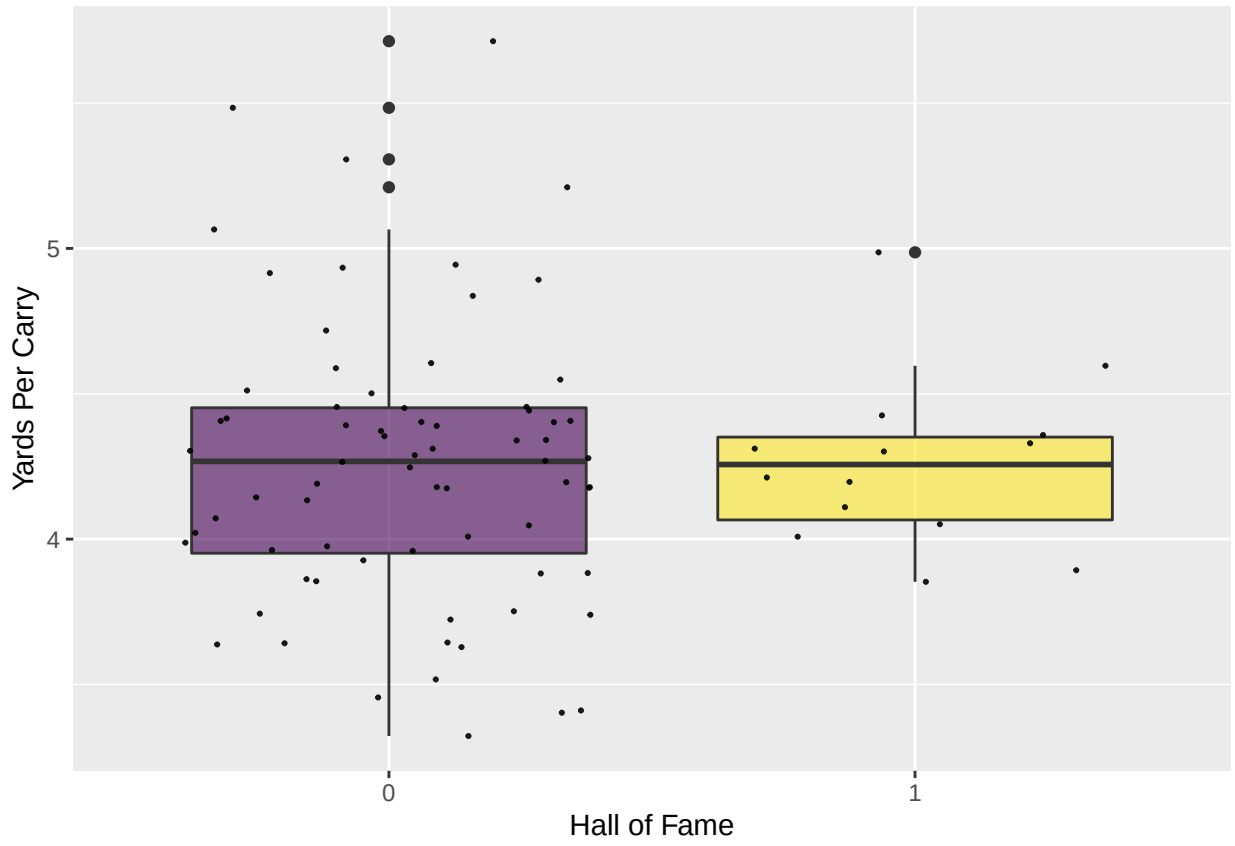

Figure 7: Boxplot of Yards per Carry for active and Hall of Fame Running Backs

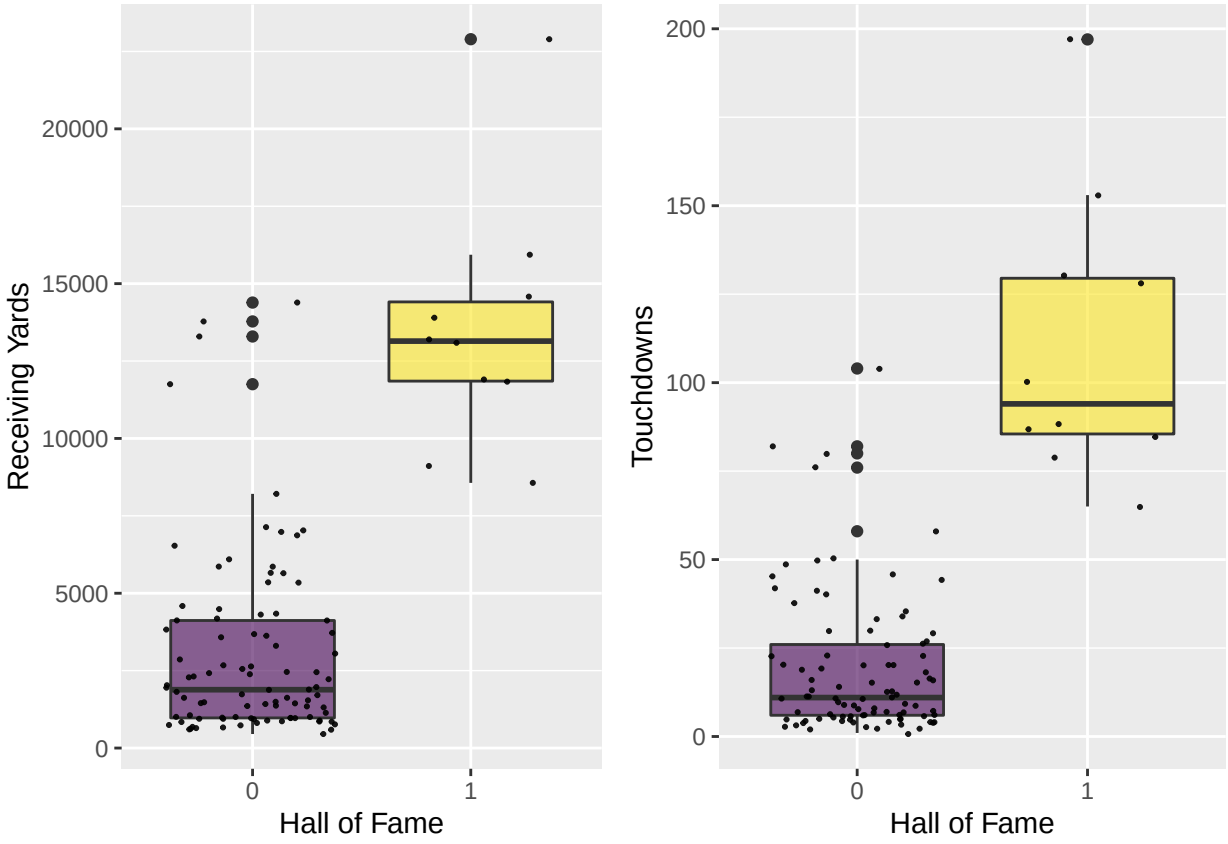

Figure 8: Boxplot of Receiving Yards and Touchdowns for current and Hall of Fame Wide Receivers

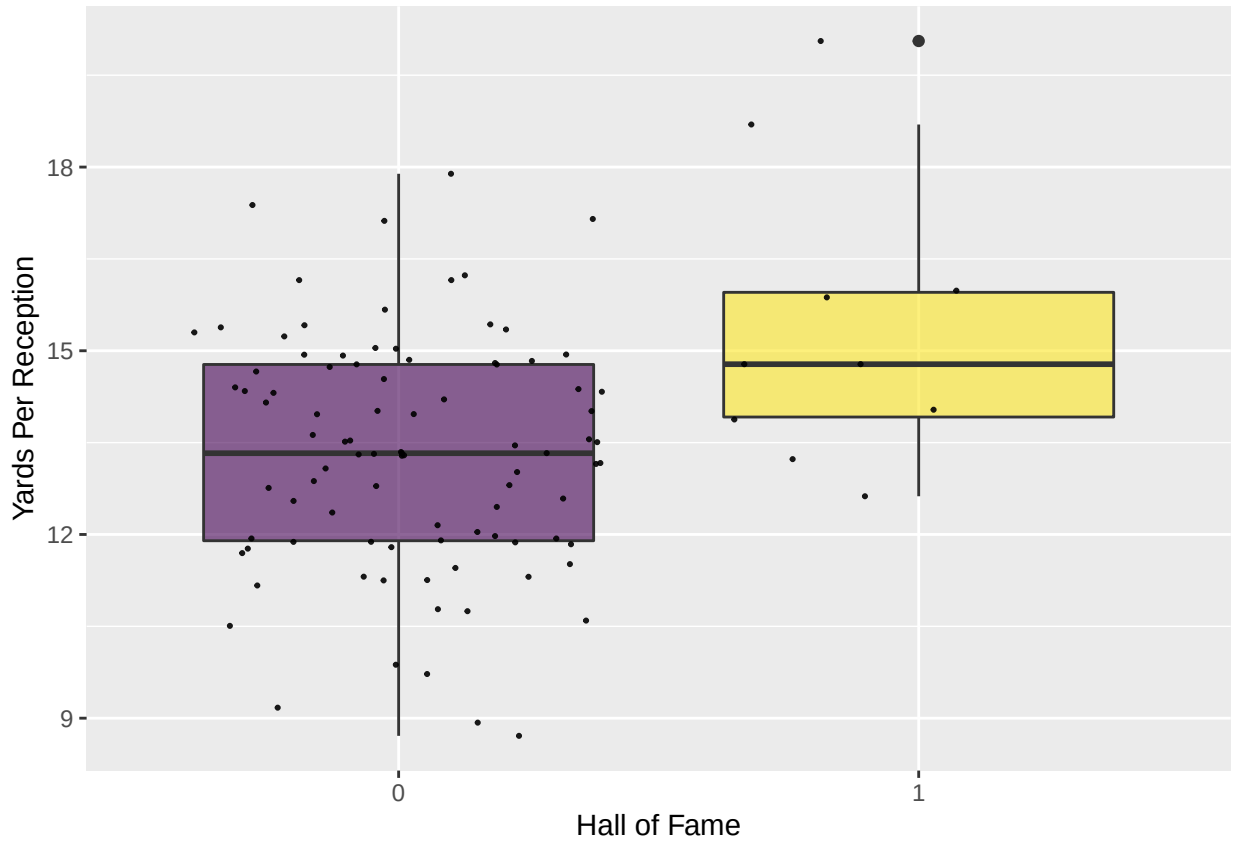

Figure 9: Boxplot of Yards per Reception for current and Hall of Fame Wide Receivers

a 0 for current players and a 1 for Hall of Famers. In the right boxplot, we see an outlier that is very far away from the rest of the data. This is Jerry Rice, arguably the best wide receiver in NFL history. The four outliers in left boxplot for both receiving yards and touchdowns are Larry Fitzgerald, Anquan Boldin, Steve Smith, and Brandon Marshall. Boldin and Smith are both retired, and Marshall is currently not signed to a team. Fitzgerald returned for his sixteenth season this fall. All four of these players have been among Hall of Fame discussion throughout their careers. Next, let's analyze the boxplot that visualizes yards per reception.

Unlike the yards per carry plot from the running back section, we see a slightly better average for the Hall of Famers in Figure 9. This shows how efficient these Hall of Fame receivers were with the ball. This can be due to statistics not in the dataset such as broken tackles or yards after catch. Thus, we should consider this variable as well when making final predictions.

By exploring all these relationships within the dataset and comparing current players to Hall of Fame players, we can move onto our methods section where we utilize statistical procedures to help us make accurate predictions on future Hall of Fame inductees.

Section 4 : Methods

Now that we have our clean datasets and have performed exploratory data analysis to determine important variables of interest, we can move on to discussing methods that will help us in making our predictions.

4.1: Clustering

The technique I will be using to identify the players that most resemble the Hall of Fame players is called **clustering**. Clustering methods involve grouping data points into **clusters**, which are subsets of the data

whose data points have similar aspects with the rest of the data within the same cluster. The goal of the clusters are to essentially separate and categorize the data into several groups. For the purpose of my report, we can use clustering to group together the players that have very similar career statistics to other players in their respective goal. We can observe the clusters that include the Hall of Fame players and see what current players are also included in that cluster.

4.1.1 : Partitional Clustering (K-means and PAM)

There are multiple types of clustering methods we can use here. The first form of clustering we can consider is **partitional clustering**. The specific type of partitional clustering we will use is **K-means clustering**. This method splits the data into k clusters that are each represented by the centroid, which is the mean of the data within the cluster. It is important to note that the **K-means** method is sensitive to outliers. Another method is K-medoids clustering, or **PAM (Partitioning Around Medoids)**, in which each cluster is represented by one of the objects in the cluster. In general, the PAM technique is less sensitive to outliers compared to k-means. Both these types of clustering are forms of what is called **hard clustering**, meaning every data point gets assigned to only one or no clusters.

The first method we will use is the K-means technique. This method involves splitting our n data points in our dataset into K disjoint subsets, or **clusters**. Every cluster that is created is represented by the geometric centroid, otherwise known as the center or means of the data points. Formulating the clusters involves using the **k-means algorithm**. The goal of using the algorithm is to minimize the within-cluster variances, which are also known as the squared Euclidean distance between each data point to the mean. The process includes two steps. Firstly, we compute the geometric centroid within the data set. For the second step, the algorithm assigns each data point with the cluster whose geometric centroid is closest to that point. The two steps are then repeated until there is no more change in the clustering assignments. The k-means algorithm is derived as follows:

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} |x_n - \mu_j|^2,$$

where S_j is a matrix of the scaled data, and μ_j is the geometric centroid, or mean of the data points in S_j [12]. To create S_j, we take every entry in our matrix dataset, subtract the mean, and then divide by the standard deviation. The result, J, represents a global minimum, as our goal is to shrink the within-cluster variances.

PAM method, or Partitioning Around Medoids, is the second method we will consider for our dataset. It is nearly identical to the k-means technique except that each cluster is represented by a singular data point rather than the means within that cluster. This involves using the **k-medoids algorithm**, where a **medoid** is the name for the data point that is located in the center of the cluster. In general, this method is less sensitive to outliers and statistical noise within the dataset. We will use the **PAM algorithm**, the most frequently used k-medoids clustering method. Just like the k-means technique, we need to decide on the number of k clusters. One difference, however, is that we will be using a **dissimilarity matrix** of the data. The computation of this matrix will show us the dissimilarities, represented by distances, between data points in the cluster under certain criteria. In our case, it will be the distance to and from the most centered point. So, our PAM algorithm is as follows:

$$J = \sum_{i=1}^{K} \sum_{x_j \in C_i} d(x_j, m_i),$$

where C_i is the dissimilarity matrix and $d(x_j, m_i)$ is the dissimilarity measurement between each $x_j \in C_i$ and the medoid m_i [15]. Ideally, we want J to minimize the sum of the dissimilarities of the data points to their closest medoid.

These formulas, of course, can be computed easily using certain R packages. We can visualize the plots produced using both of these techniques. The packages associated with these forms of clustering are `stats` [14], `cluster` [13], and `factoextra` [3]. Starting off with k-means, the first step is always to determine the optimal number of clusters to use for our dataset. We will begin with looking at the quarterback dataset.

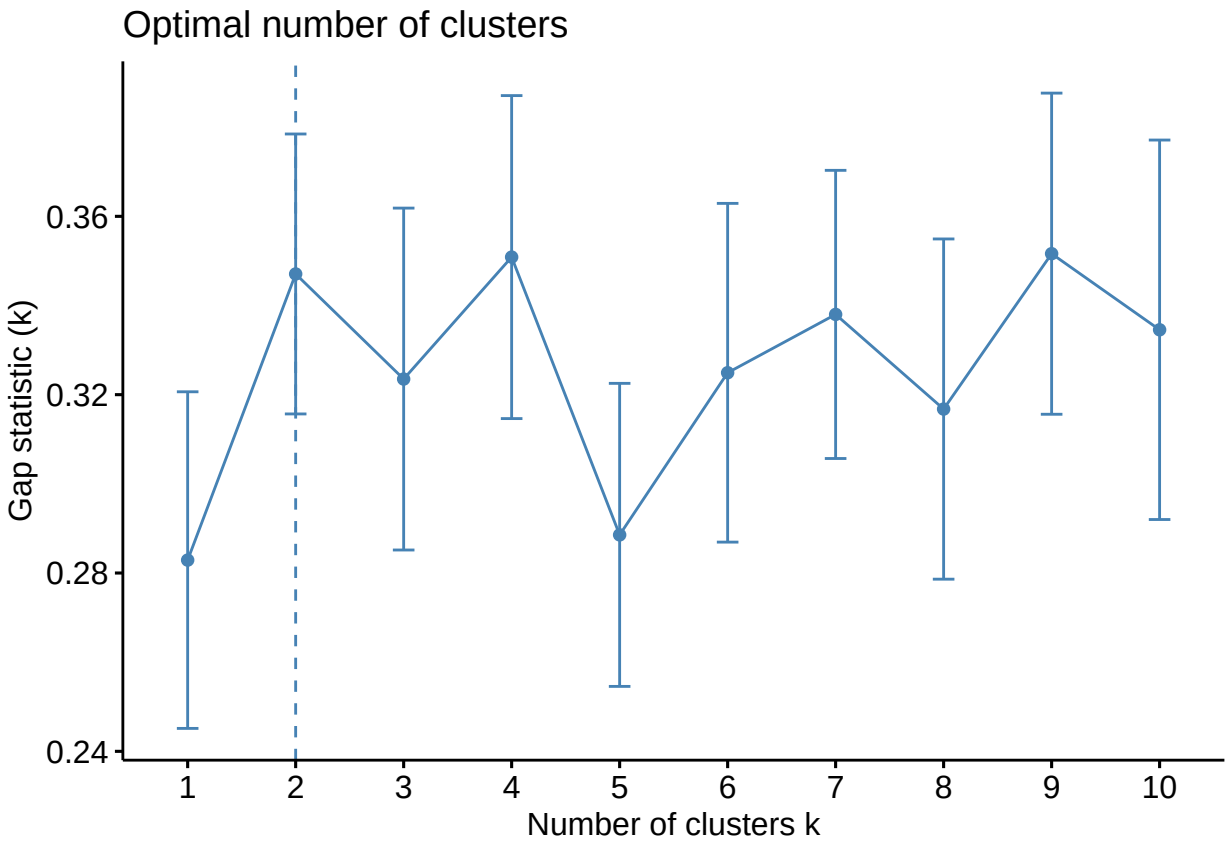

Figure 10: Optimal Number of Clusters for Quarterbacks (K-means)

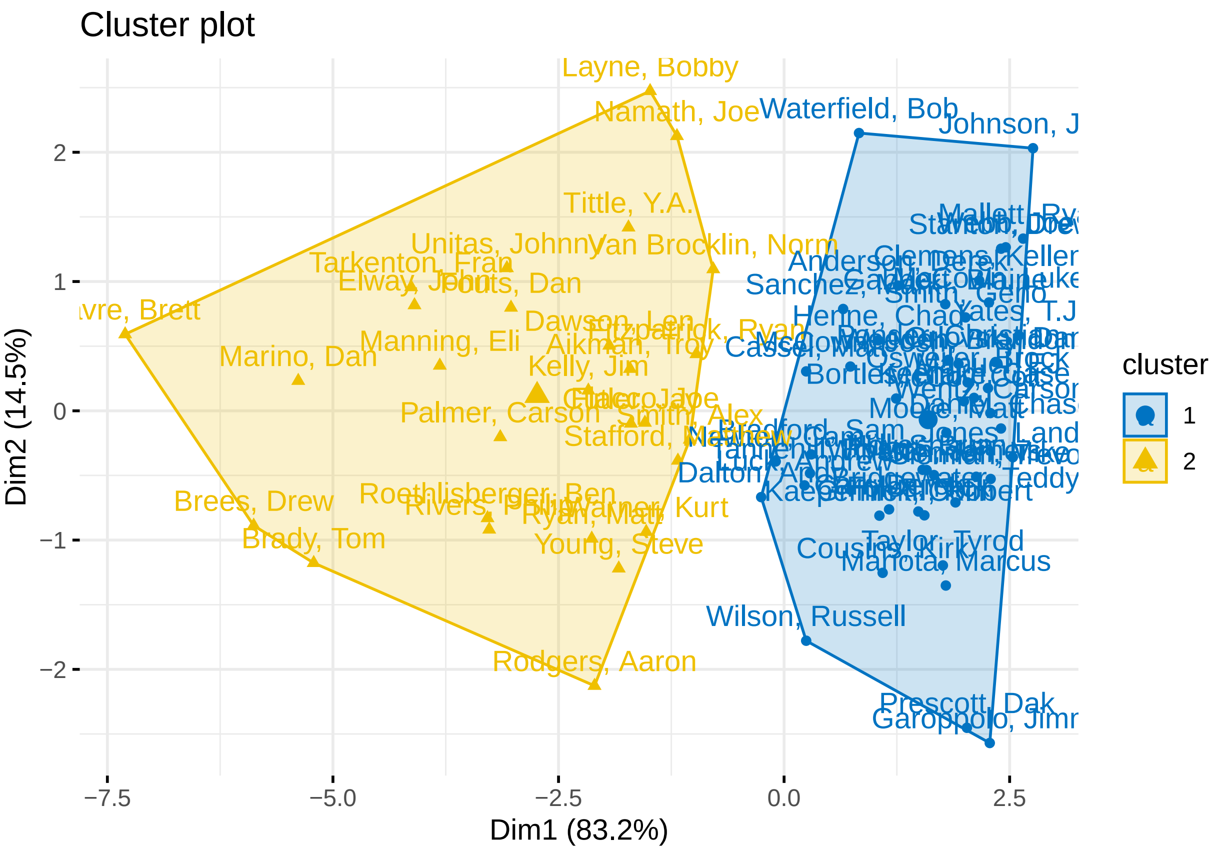

Figure 11: K-means Clustering Plot for Quarterbacks

Quarterbacks

In Figure 10, using the command `fviz_nbclust` [3] takes our dataset into consideration and determines the optimal number of k clusters to separate our data into. We can see that $k = 2$ clusters will be the most ideal, which makes sense in our context since we want to separate the players in two groups with one of them consisting of Hall of Fame players. Now with this information, we can move on with computing and visualizing clusters using the k-means technique. We use the `kmeans` command [14] with our dataset and number of clusters to create a variable to be plotted using the `fviz_cluster` command [3].

In Figure 11, we can see the two clusters created with the separated data points that are represented by the geometric centroid. The X and Y axes are represented by the **principal component analysis (PCA)** algorithm. The first and second principal component dimensions help plot the data. The PCA algorithm operates on all the variables in the dataset and creates two new variables, Dim1 and Dim2. These variables on the axes represent the originally inputted variables as a whole and also serve as the indicated amount of variation in the dataset. The yellow cluster on the left contains most of the Hall of Fame players such as Kurt Warner, Steve Young, Brett Favre, Dan Marino, and more. We also see the current players that are most similar in career numbers to these Hall of Famers. The blue cluster represents the other players that do not meet the criteria set by the Hall of Famers. Interesting enough, Bob Waterfield is included in the blue cluster despite being a Hall of Fame inductee. This is likely because he was among the first few inductees in the 1960s after leading his team to two NFL championships, and the decision was less weighed on statistics and more on the fact he led his team to victory. Since the plot is pretty cluttered and not all the names are legible, we can look who exactly landed in the first and second cluster by creating a table using `xtable` [7]:

Names	Cluster	Names	Cluster
Aikman, Troy	2	Hoyer, Brian	1
Dawson, Len	2	Johnson, Josh	1
Elway, John	2	Jones, Landry	1
Favre, Brett	2	Kaepernick, Colin	1
Fouts, Dan	2	Keenum, Case	1
Kelly, Jim	2	Luck, Andrew	1
Layne, Bobby	2	Mallett, Ryan	1
Marino, Dan	2	Manning, Eli	2
Namath, Joe	2	Manuel, EJ	1
Tarkenton, Fran	2	Mariota, Marcus	1
Tittle, Y.A.	2	McCown, Josh	1
Unitas, Johnny	2	McCown, Luke	1
Van Brocklin, Norm	2	McCoy, Colt	1
Warner, Kurt	2	Moore, Matt	1
Waterfield, Bob	1	Newton, Cam	1
Young, Steve	2	Orlovsky, Dan	1
Anderson, Derek	1	Osweiler, Brock	1
Bortles, Blake	1	Palmer, Carson	2
Bradford, Sam	1	Ponder, Christian	1
Brady, Tom	2	Prescott, Dak	1
Brees, Drew	2	Rivers, Philip	2
Bridgewater, Teddy	1	Rodgers, Aaron	2
Carr, Derek	1	Roethlisberger, Ben	2
Cassel, Matt	1	Ryan, Matt	2
Clemens, Kellen	1	Sanchez, Mark	1
Cousins, Kirk	1	Siemian, Trevor	1
Cutler, Jay	2	Smith, Alex	2
Dalton, Andy	1	Smith, Geno	1
Daniel, Chase	1	Stafford, Matthew	2
Fitzpatrick, Ryan	2	Stanton, Drew	1
Flacco, Joe	2	Tannehill, Ryan	1
Foles, Nick	1	Taylor, Tyrod	1
Gabbert, Blaine	1	Webb, Joe	1
Garoppolo, Jimmy	1	Weeden, Brandon	1
Glennon, Mike	1	Wentz, Carson	1
Griffin III, Robert	1	Wilson, Russell	1
Henne, Chad	1	Winston, Jameis	1
Hill, Shaun	1	Yates, T.J.	1

Apart from Waterfield, all of the Hall of Fame quarterbacks ended up in the second cluster. This method also shows us the active quarterbacks that ended up in the second cluster whose career numbers most closely resemble those of the Hall of Fame players. These names include Tom Brady, Drew Brees, Jay Cutler, Joe Flacco, Ryan Fitzpatrick, Eli Manning, Carson Palmer, Philip Rivers, Aaron Rodgers, Matt Ryan, Ben Roethlisberger, Matthew Stafford, and Alex Smith. From what I've seen on sources like ESPN, I know some of the names, like Alex Smith and Jay Cutler, are not considered by the media to be future Hall of Famers. This is supported by their career statistics not coming close to someone like Drew Brees or Aaron Rodgers. Additionally, in today's NFL, Cutler and Smith aren't even starting quarterbacks for any team as of now.

Since we used the K-means method, the results should be sensitive to outliers. We can try the PAM clustering technique next and see if we have any additional or left out players in the clusters.

The number of clusters will again be $k = 2$ based on Figure 12. Our next step is to calculate the dissimilarity matrix to be used for algorithm. We can do this using the `daisy` command in the `cluster` package [13].

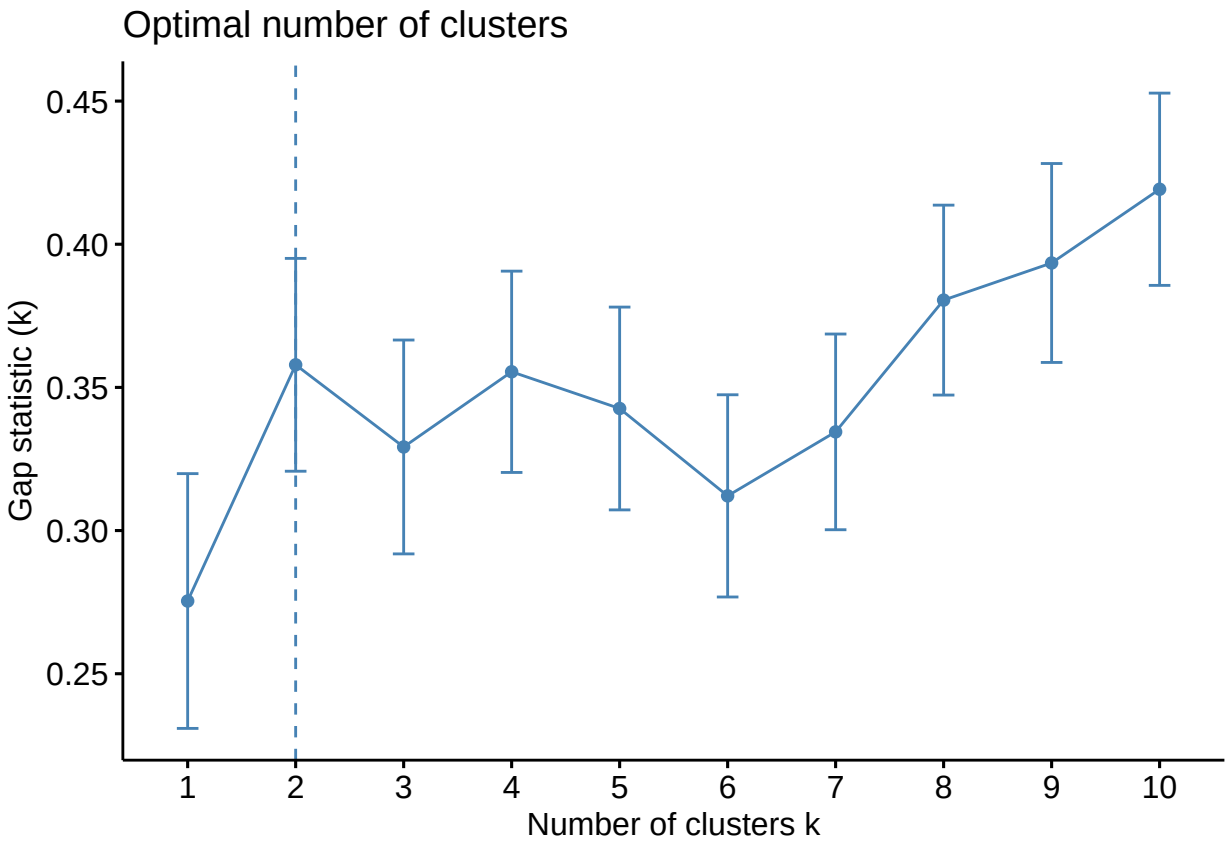

Figure 12: Optimal Number of Clusters for Quarterbacks (PAM)

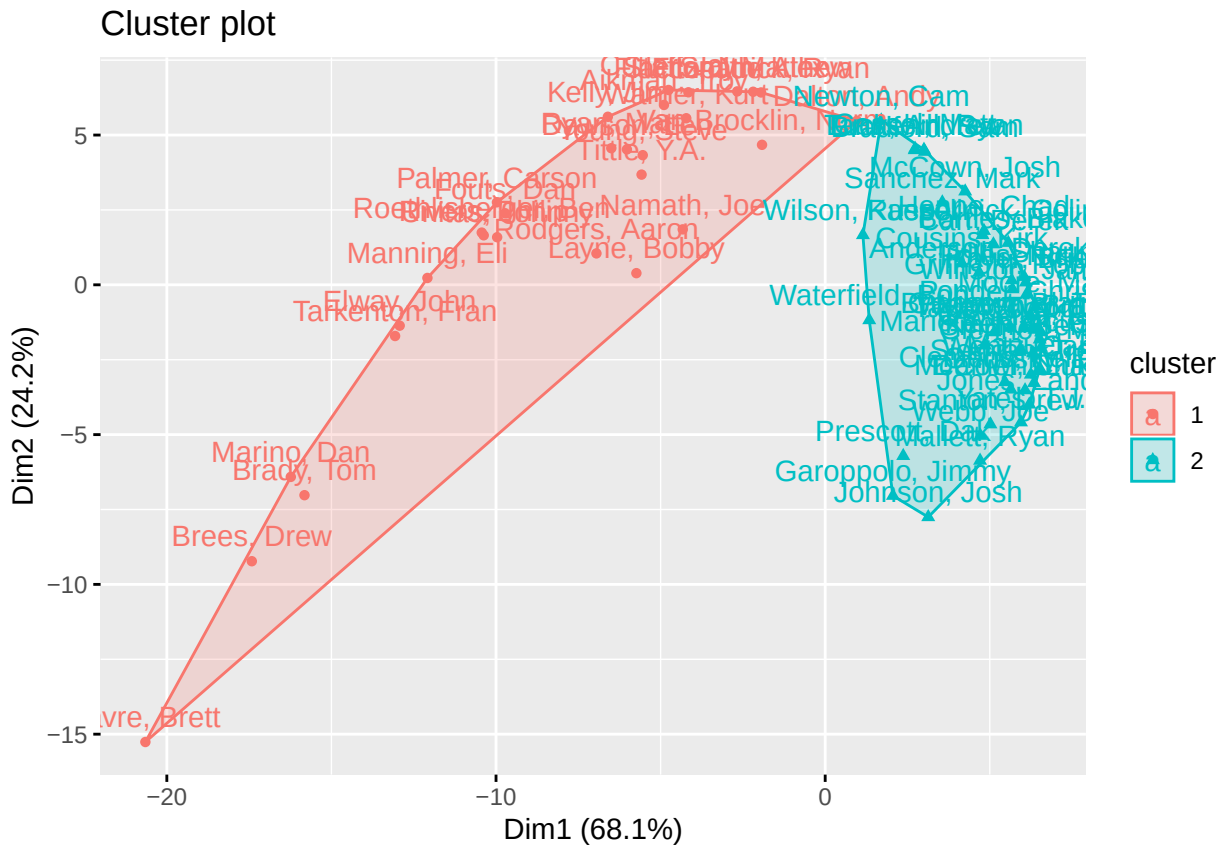

Figure 13: PAM Clustering Plot for Quarterbacks

Using this matrix, we can now use the `pam` command [13] and then visualize the result with `fviz_cluster` [3].

In Figure 13, we see that the shapes of the clusters have changed drastically. Again, the axes are represented by the principal dimensions. Since the PAM algorithm is less sensitive to noise and outliers, there is a possibility that this algorithm includes some more players in our Hall of Fame cluster than the K-means technique had. Again, the names are very cluttered, so let's look at the output of the PAM clustering using `xtable` [7]:

Names	Cluster	Names	Cluster
Aikman, Troy	1	Hoyer, Brian	2
Dawson, Len	1	Johnson, Josh	2
Elway, John	1	Jones, Landry	2
Favre, Brett	1	Kaepernick, Colin	2
Fouts, Dan	1	Keenum, Case	2
Kelly, Jim	1	Luck, Andrew	2
Layne, Bobby	1	Mallett, Ryan	2
Marino, Dan	1	Manning, Eli	1
Namath, Joe	1	Manuel, EJ	2
Tarkenton, Fran	1	Mariota, Marcus	2
Tittle, Y.A.	1	McCown, Josh	2
Unitas, Johnny	1	McCown, Luke	2
Van Brocklin, Norm	1	McCoy, Colt	2
Warner, Kurt	1	Moore, Matt	2
Waterfield, Bob	2	Newton, Cam	2
Young, Steve	1	Orlovsky, Dan	2
Anderson, Derek	2	Osweiler, Brock	2
Bortles, Blake	2	Palmer, Carson	1
Bradford, Sam	2	Ponder, Christian	2
Brady, Tom	1	Prescott, Dak	2
Brees, Drew	1	Rivers, Philip	1
Bridgewater, Teddy	2	Rodgers, Aaron	1
Carr, Derek	2	Roethlisberger, Ben	1
Cassel, Matt	2	Ryan, Matt	1
Clemens, Kellen	2	Sanchez, Mark	2
Cousins, Kirk	2	Siemian, Trevor	2
Cutler, Jay	1	Smith, Alex	1
Dalton, Andy	1	Smith, Geno	2
Daniel, Chase	2	Stafford, Matthew	1
Fitzpatrick, Ryan	1	Stanton, Drew	2
Flacco, Joe	1	Tannehill, Ryan	2
Foles, Nick	2	Taylor, Tyrod	2
Gabbert, Blaine	2	Webb, Joe	2
Garoppolo, Jimmy	2	Weeden, Brandon	2
Glennon, Mike	2	Wentz, Carson	2
Griffin III, Robert	2	Wilson, Russell	2
Henne, Chad	2	Winston, Jameis	2
Hill, Shaun	2	Yates, T.J.	2

This time, the cluster that includes the Hall of Fame players is now indicated as cluster 1. Our output tells us that only one new player has been added using the PAM technique, and that is Andy Dalton. Dalton is another quarterback who has barely gotten any praise from the media. While he has had a couple decent years and been selected to a couple Pro Bowls, his Career Passer Rating does not exceed 90, and his touchdowns to interceptions ratio is not as impressive as some of the other quarterbacks in the cluster. However, it is worth noting this change using the PAM method, as this technique is generally less sensitive to outliers. We can explore other methods of clustering to see if we get the same results.

Running backs

Next, we will use these two methods of clustering with the running backs. We use the same packages and commands used in the previous section with the quarterbacks. Starting with the K-means technique, we determine the ideal number of clusters.

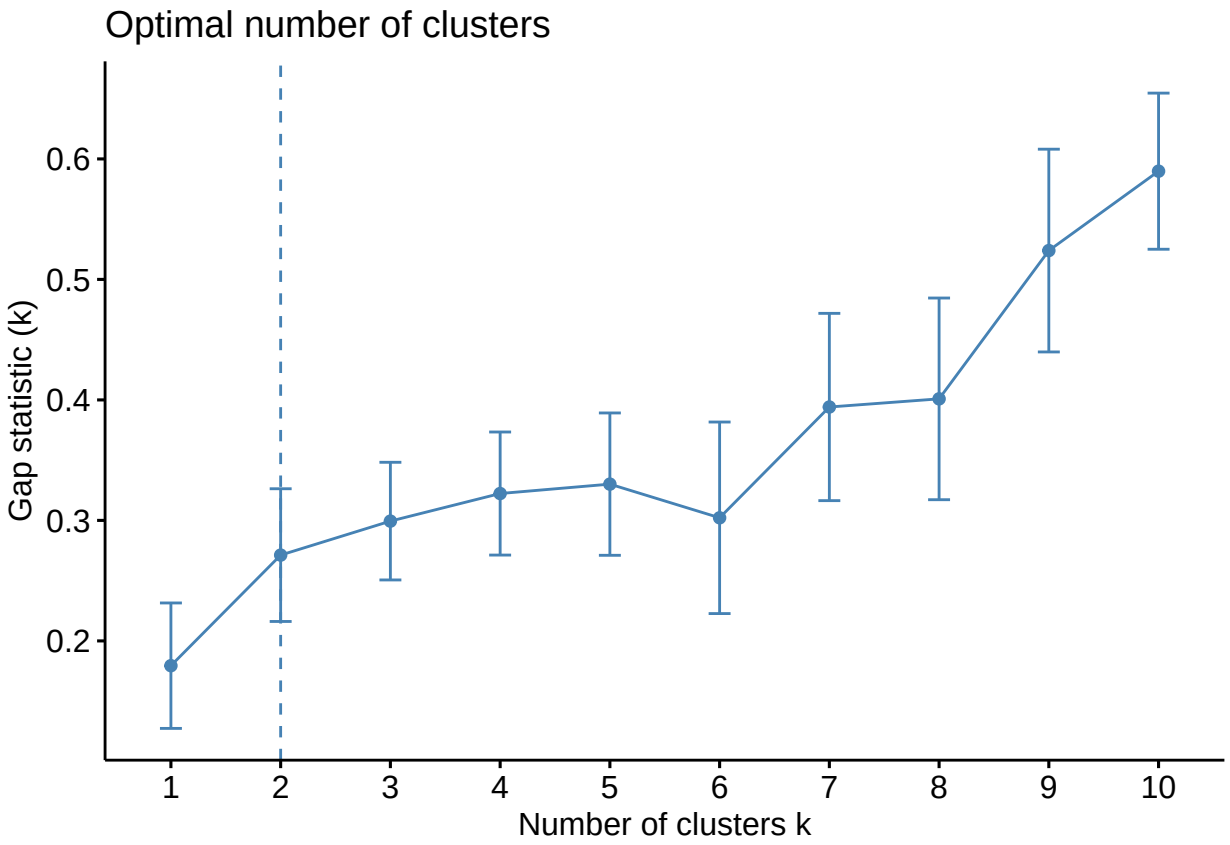

Figure 14: Optimal Number of Clusters for Running Backs (K-means)

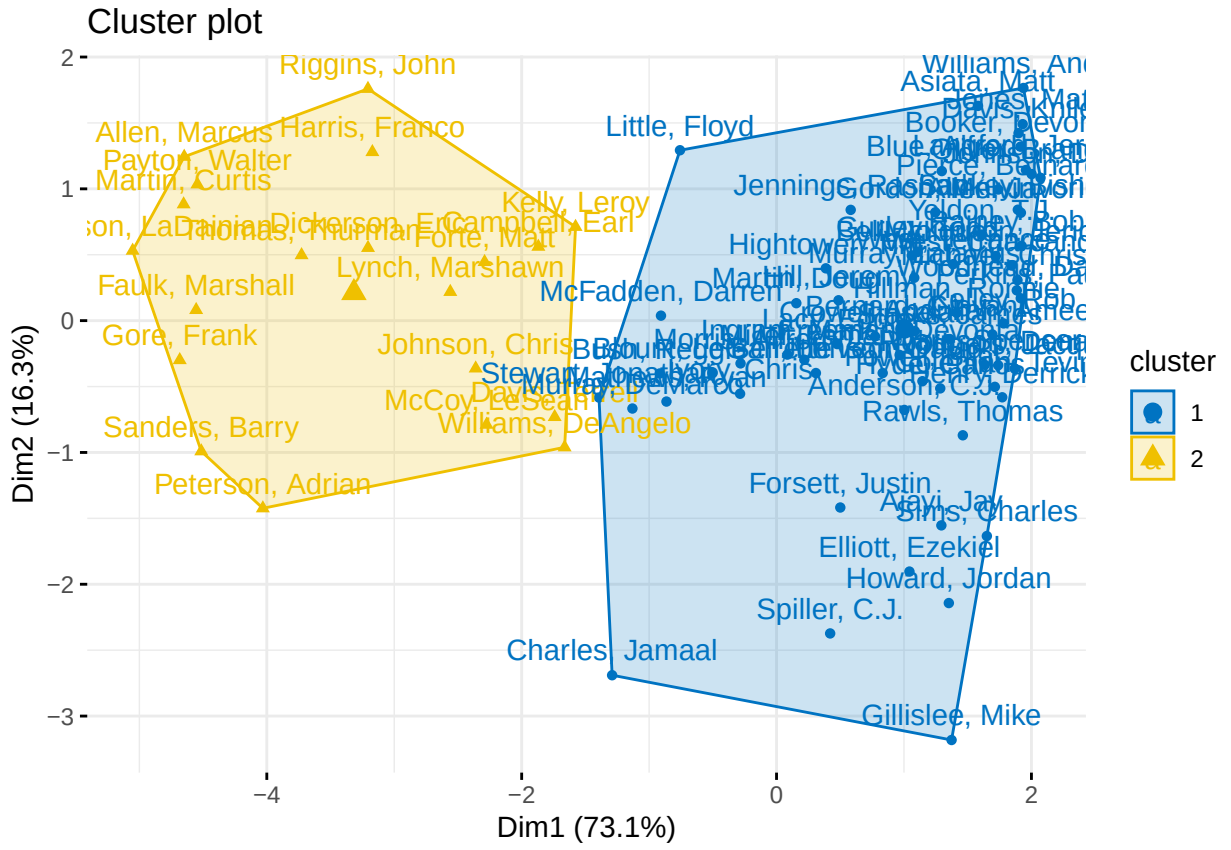

Figure 15: K-means Clustering Plot for Running Backs

Figure 14 tells us that $k = 2$ clusters is optimal for our dataset. Again, this makes sense since we're comparing Hall of Fame players to the current players. Now we can plot the clusters with `fviz_cluster` [3].

The results of our k-means clustering is displayed in 15. In the left cluster, we see most of the Hall of Fame running back, but it is hard to see the majority of the names in the right cluster. We can use `xtable` [7] to observe who exactly landed in each cluster.

Name	Cluster	Name	Cluster
Allen, Marcus	2	Hyde, Carlos	1
Campbell, Earl	2	Ingram, Mark	1
Davis, Terrell	2	Ivory, Chris	1
Dickerson, Eric	2	Jennings, Rashad	1
Faulk, Marshall	2	Johnson, Chris	2
Harris, Franco	2	Johnson, David	1
Kelly, Leroy	2	Johnson, Duke	1
Little, Floyd	1	Jones, Matt	1
Martin, Curtis	2	Kelley, Rob	1
Payton, Walter	2	Lacy, Eddie	1
Riggins, John	2	Langford, Jeremy	1
Sanders, Barry	2	Lynch, Marshawn	2
Thomas, Thurman	2	Martin, Doug	1
Tomlinson, LaDainian	2	Mathews, Ryan	1
Abdullah, Ameer	1	McCoy, LeSean	2
Ajayi, Jay	1	McFadden, Darren	1
Allen, Javorius	1	McKinnon, Jerick	1
Anderson, C.J.	1	Michael, Christine	1
Asiata, Matt	1	Miller, Lamar	1
Bell, Joique	1	Morris, Alfred	1
Bell, Le'Veon	1	Murray, DeMarco	1
Bernard, Giovani	1	Murray, Latavius	1
Blount, LeGarrette	1	Oliver, Branden	1
Blue, Alfred	1	Perkins, Paul	1
Booker, Devontae	1	Peterson, Adrian	2
Bush, Reggie	1	Pierce, Bernard	1
Charles, Jamaal	1	Powell, Bilal	1
Coleman, Tevin	1	Rainey, Bobby	1
Crowell, Isaiah	1	Rawls, Thomas	1
Davis, Knile	1	Robinson, Denard	1
Elliott, Ezekiel	1	Rodgers, Jacquizz	1
Forsett, Justin	1	Sankey, Bishop	1
Forte, Matt	2	Sims, Charles	1
Freeman, Devonta	1	Spiller, C.J.	1
Gillislee, Mike	1	Starks, James	1
Gordon, Melvin	1	Stewart, Jonathan	1
Gore, Frank	2	Ware, Spencer	1
Gurley, Todd	1	West, Charcandrick	1
Henry, Derrick	1	West, Terrance	1
Hightower, Tim	1	Williams, Andre	1
Hill, Jeremy	1	Williams, DeAngelo	2
Hillman, Ronnie	1	Woodhead, Danny	1
Howard, Jordan	1	Yeldon, T.J.	1

Similar to Waterfield from the quarterback section, we see that one of the Hall of Famers, Floyd Little, is not included in the cluster with the rest of the Hall of Fame players. When looking at the dataset, he has the lowest yardage and touchdown totals among the other players, so the clustering method paired him with the other cluster. As for other players, we don't see many players that were included in the cluster with Hall of Fame players. However, the small number of players included in cluster 2 is pretty accurate in my opinion based on their stats. Because of potential outliers and noise, we now switch to the PAM clustering technique to see if our results will be any different.

In Figure 16, for the first time we see $k = 3$ clusters using the PAM technique. We can use the `daisy` [13]

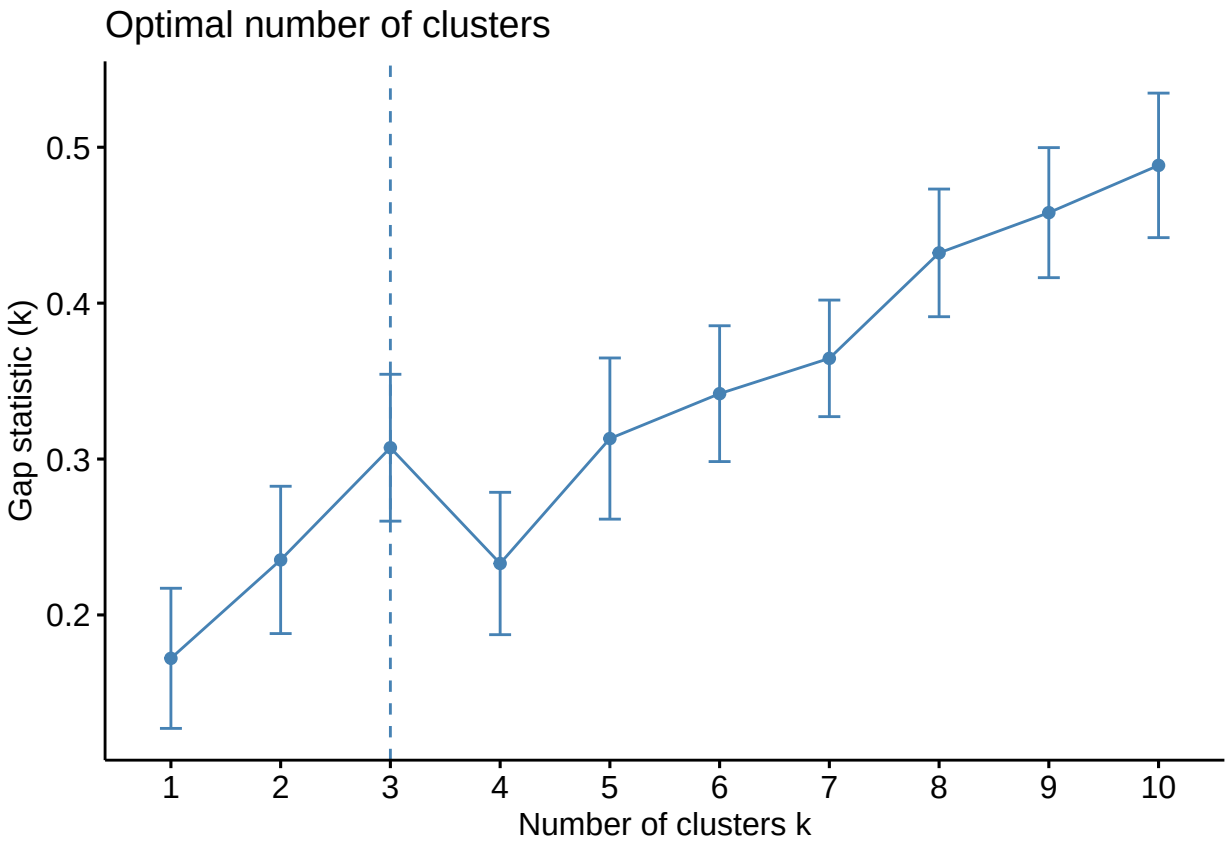

Figure 16: Optimal Number of Clusters for Running Backs (PAM)

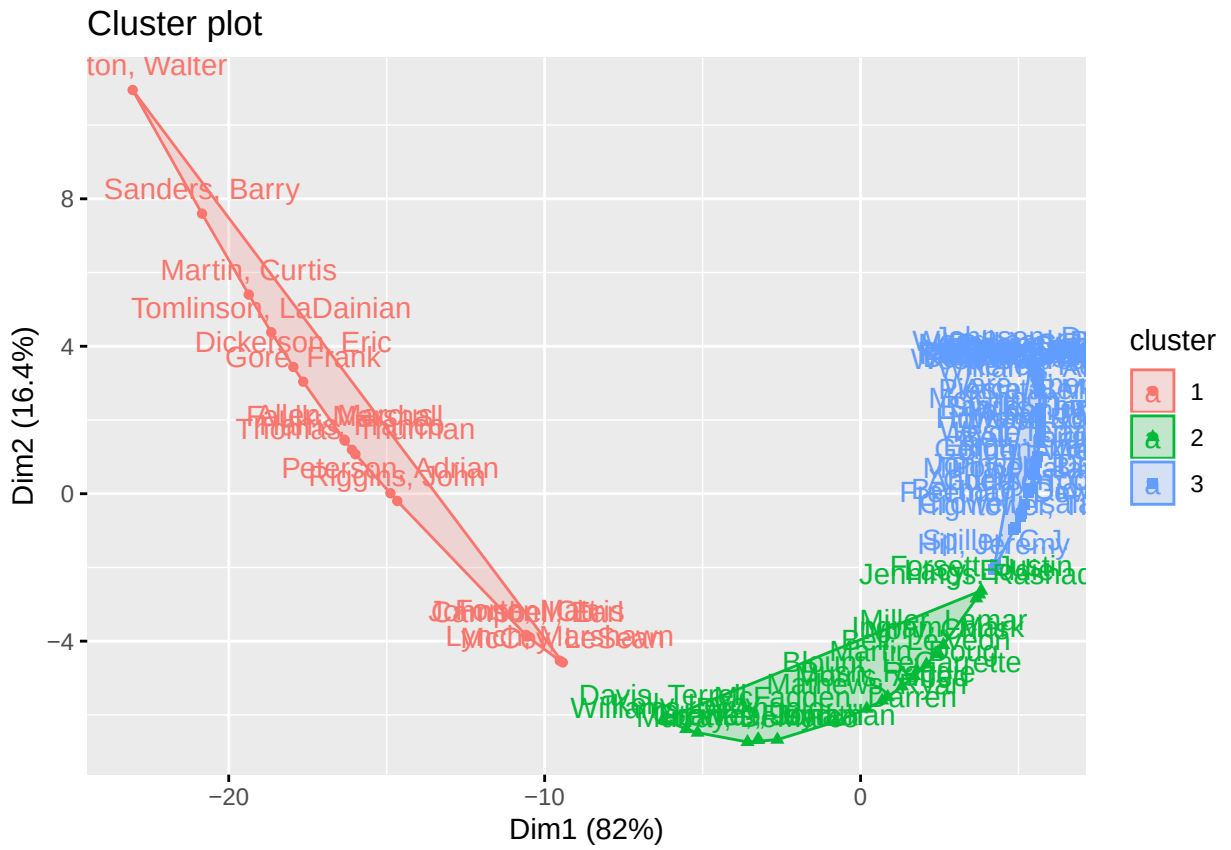

Figure 17: PAM Clustering Plot for Running Backs

command to create the dissimilarity matrix followed by the `pam` [13] and `fviz_cluster` [3] commands to visualize the 3 clusters.

Clearly, Figure 17 is very hard to observe, so we will use `xtable` [7] again to see who ended up in each cluster.

Name	Cluster	Name	Cluster
Allen, Marcus	1	Hyde, Carlos	3
Campbell, Earl	1	Ingram, Mark	2
Davis, Terrell	2	Ivory, Chris	2
Dickerson, Eric	1	Jennings, Rashad	2
Faulk, Marshall	1	Johnson, Chris	1
Harris, Franco	1	Johnson, David	3
Kelly, Leroy	2	Johnson, Duke	3
Little, Floyd	2	Jones, Matt	3
Martin, Curtis	1	Kelley, Rob	3
Payton, Walter	1	Lacy, Eddie	2
Riggins, John	1	Langford, Jeremy	3
Sanders, Barry	1	Lynch, Marshawn	1
Thomas, Thurman	1	Martin, Doug	2
Tomlinson, LaDainian	1	Mathews, Ryan	2
Abdullah, Ameer	3	McCoy, LeSean	1
Ajayi, Jay	3	McFadden, Darren	2
Allen, Javorius	3	McKinnon, Jerick	3
Anderson, C.J.	3	Michael, Christine	3
Asiata, Matt	3	Miller, Lamar	2
Bell, Joique	3	Morris, Alfred	2
Bell, Le'Veon	2	Murray, DeMarco	2
Bernard, Giovani	3	Murray, Latavius	3
Blount, LeGarrette	2	Oliver, Branden	3
Blue, Alfred	3	Perkins, Paul	3
Booker, Devontae	3	Peterson, Adrian	1
Bush, Reggie	2	Pierce, Bernard	3
Charles, Jamaal	2	Powell, Bilal	3
Coleman, Tevin	3	Rainey, Bobby	3
Crowell, Isaiah	3	Rawls, Thomas	3
Davis, Knile	3	Robinson, Denard	3
Elliott, Ezekiel	3	Rodgers, Jacquizz	3
Forsett, Justin	2	Sankey, Bishop	3
Forte, Matt	1	Sims, Charles	3
Freeman, Devonta	3	Spiller, C.J.	3
Gillislee, Mike	3	Starks, James	3
Gordon, Melvin	3	Stewart, Jonathan	2
Gore, Frank	1	Ware, Spencer	3
Gurley, Todd	3	West, Charcandrick	3
Henry, Derrick	3	West, Terrance	3
Hightower, Tim	3	Williams, Andre	3
Hill, Jeremy	3	Williams, DeAngelo	2
Hillman, Ronnie	3	Woodhead, Danny	3
Howard, Jordan	3	Yeldon, T.J.	3

From our table, we can interpret our 3 clusters differently. The first cluster includes the majority of the Hall of Famers and the current players that were included in the Hall of Fame cluster using K-means. The second cluster includes the rest of the Hall of Fame players as well as several other current players. The third cluster includes the rest of the players from the dataset. From this, I can say that those players includes in cluster 2 may not be on the level of the Hall of Fame players at this point in their careers, but if they keep up with their current yearly averages, they may end up being considered for the Hall of Fame by the end of their careers. For example, Le'Veon Bell was placed in cluster 2, and he has been regarded as one of the best running backs in the past decade. He recently got left the Steelers to join the Jets. The new offense

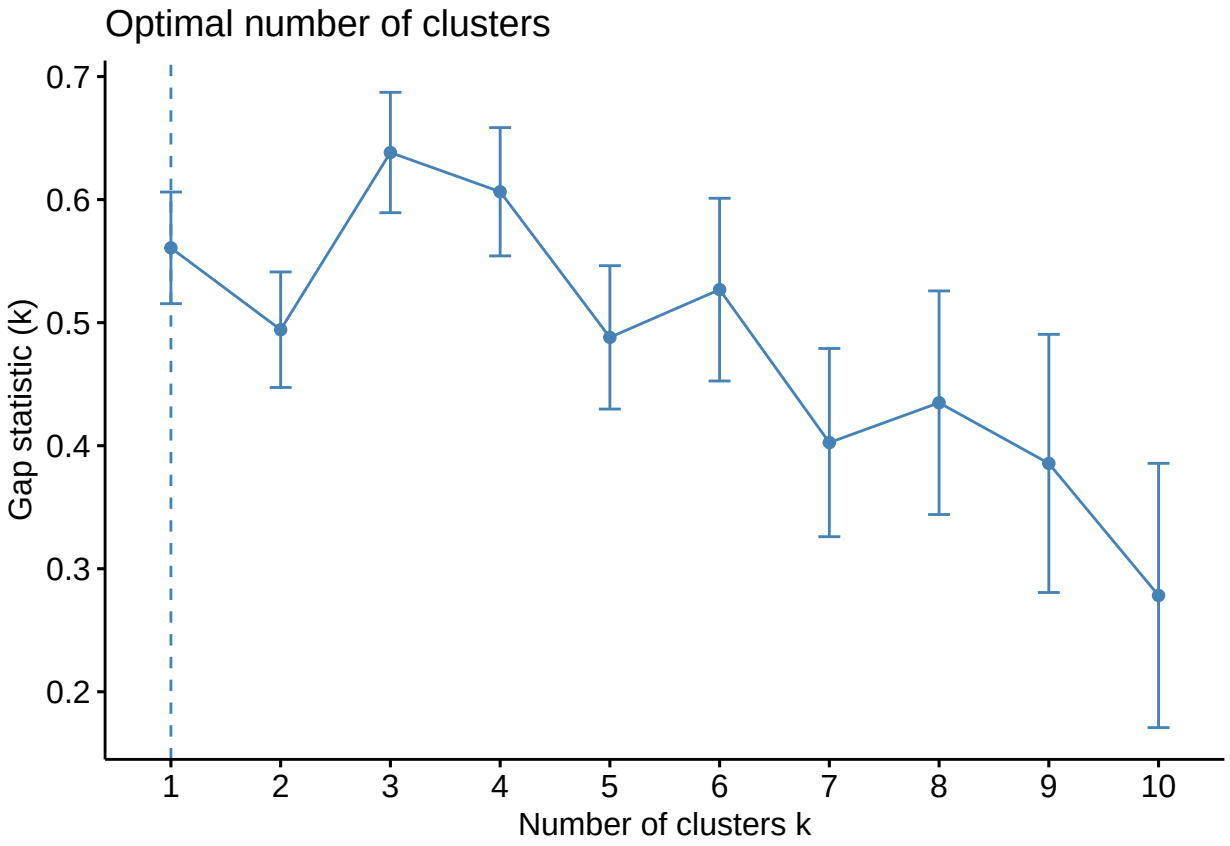

Figure 18: Optimal Number of Clusters for Wide Receivers(K-means)

he is playing in is not as dominant as the Steelers, as the Jets usually finish in the bottom of their division. However, if he can produce similar numbers than he did in his prime with the Steelers, there is a high chance he could still be inducted into the Hall of Fame. Since we have the 3 clusters, we can use fuzzy clustering in the next section to observe the exact probabilities for each player to be placed in the Hall of Fame clusters.

Wide Receivers

Lastly, we have our wide receiver dataset. The same packages and commands are used throughout this section. We will start with the k-means method and check the optimal number of clusters to use.

For the first time, Figure 18 tells us to use only one cluster for the wide receivers. For my report, I will be disregarding this output because then there would be no way to distinguish active and Hall of Fame players. Additionally, as we saw from the EDA section, there are clear differences in statistics between the two categories of players. For these reasons, I will stick to using $k = 2$ clusters when using the `fviz_cluster` command [3].

Based on 19, the blue cluster includes the majority of the Hall of Famer receivers plus some of the players I discussed have had amazing careers like Larry Fitzgerald and Steve Smith. Let's look at the `xtable` [7] output to see who ended up in the yellow cluster that has a lot of cluttered names.

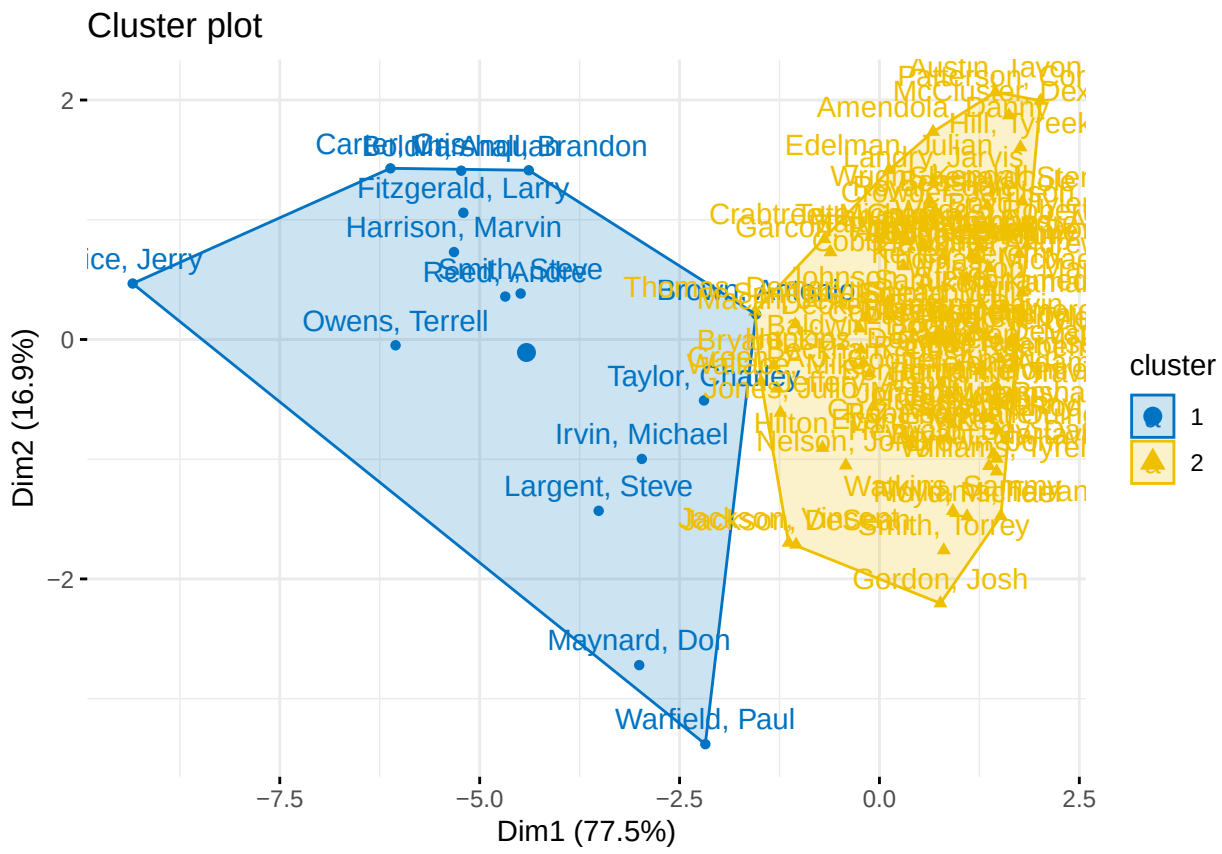

Figure 19: K-means Clustering Plot for Wide Receivers

Name	Cluster	Name	Cluster
Carter, Cris	1	Hilton, T.Y.	2
Harrison, Marvin	1	Hopkins, DeAndre	2
Irvin, Michael	1	Humphries, Adam	2
Largent, Steve	1	Hurns, Allen	2
Maynard, Don	1	Inman, Dontrelle	2
Owens, Terrell	1	Jackson, DeSean	2
Reed, Andre	1	Jackson, Vincent	2
Rice, Jerry	1	Jeffery, Alshon	2
Taylor, Charley	1	Johnson, Steve	2
Warfield, Paul	1	Jones, Julio	2
Adams, Davante	2	Jones, Marvin	2
Aiken, Kamar	2	Kerley, Jeremy	2
Allen, Keenan	2	LaFell, Brandon	2
Amendola, Danny	2	Landry, Jarvis	2
Austin, Tavon	2	Lee, Marqise	2
Baldwin, Doug	2	Lockett, Tyler	2
Beasley, Cole	2	Maclin, Jeremy	2
Beckham, Odell	2	Marshall, Brandon	1
Benjamin, Kelvin	2	Matthews, Jordan	2
Benjamin, Travis	2	Matthews, Rishard	2
Blackmon, Justin	2	McCluster, Dexter	2
Boldin, Anquan	1	Meredith, Cameron	2
Boyd, Tyler	2	Moncrief, Donte	2
Britt, Kenny	2	Nelson, Jordy	2
Brown, Antonio	2	Parker, DeVante	2
Brown, John	2	Patterson, Cordarrelle	2
Bryant, Dez	2	Pryor, Terrelle	2
Bryant, Martavis	2	Randle, Rueben	2
Cobb, Randall	2	Roberts, Andre	2
Cooks, Brandin	2	Robinson, Allen	2
Cooper, Amari	2	Royal, Eddie	2
Crabtree, Michael	2	Sanders, Emmanuel	2
Crowder, Jamison	2	Sanu, Mohamed	2
Cruz, Victor	2	Shepard, Sterling	2
Decker, Eric	2	Shorts, Cecil	2
Diggs, Stefon	2	Smith, Steve	1
Douglas, Harry	2	Smith, Torrey	2
Edelman, Julian	2	Snead, Willie	2
Enunwa, Quincy	2	Stills, Kenny	2
Evans, Mike	2	Streater, Rod	2
Fitzgerald, Larry	1	Tate, Golden	2
Floyd, Michael	2	Thielen, Adam	2
Garcon, Pierre	2	Thomas, Demaryius	2
Ginn, Ted	2	Thomas, Michael	2
Gordon, Josh	2	Wallace, Mike	2
Green, A.J.	2	Watkins, Sammy	2
Harvin, Percy	2	Wheaton, Markus	2
Hawkins, Andrew	2	Williams, Terrance	2
Hester, Devin	2	Williams, Tyrell	2
Heyward-Bey, Darrius	2	Woods, Robert	2
Hill, Tyreek	2	Wright, Kendall	2

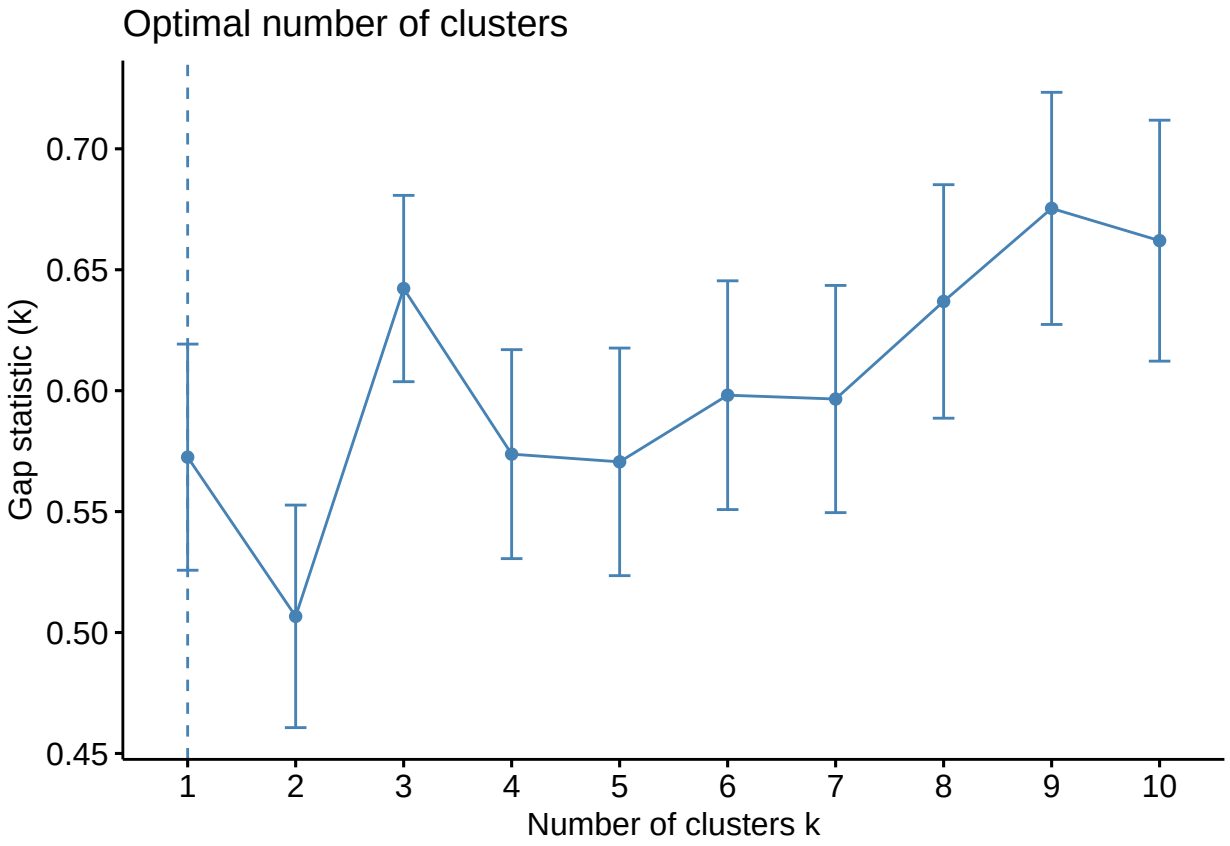

Figure 20: Optimal Number of Clusters for Wide Receivers (PAM)

As predicted, all of the Hall of Fame players were included in cluster 1, as well as the four receivers that were represented as outliers in the boxplots from the EDA section. However, from personal football knowledge, there are a good number of receivers included in cluster 2 that I believe could end up in the Hall of Fame by the end of their careers based on their reputation today. For instance, DeAndre Hopkins is said to have the "best hands" by the sports media world and makes amazing catches weekly. The reason behind his cluster placement may be due to his young age and the fact he still has a long way in his career to go. The same thing can be said about Odell Beckham, who made arguably the "best catch" ever seen a couple of seasons ago with the New York Giants. These will be factors I will consider in my final predictions. For now, we will now use the PAM clustering method to see if our clusters change at all.

Based on Figure 20, the PAM method also tells us to use one cluster. However, we will again use $k = 2$ clusters when visualizing. We calculate the dissimilarity matrix with the `daisy` command [13] and then plot the clusters.

Upon initial observation, it looks like the same names appear in cluster 1 in Figure 21. We can use another xtable [7] to confirm this assumption.

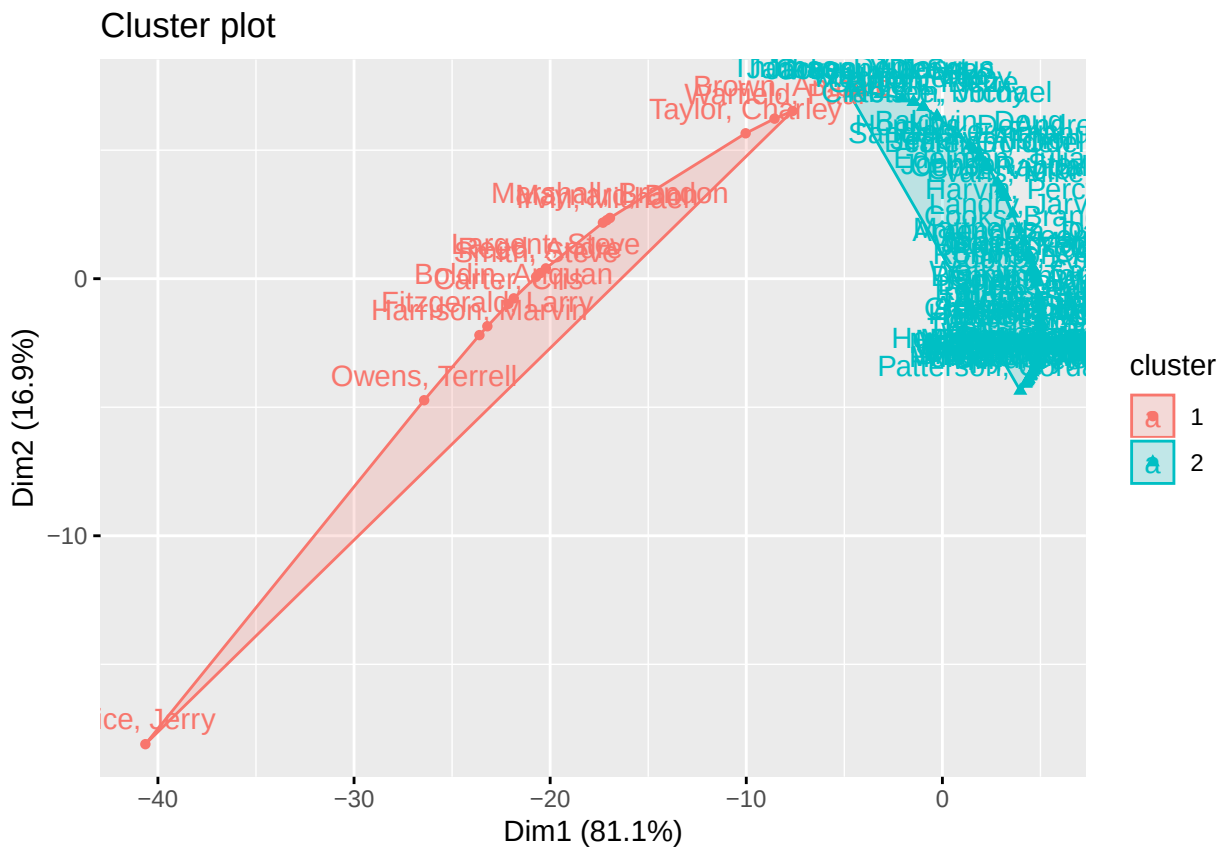

Figure 21: PAM Clustering Plot for Wide Receivers

Name	Cluster	Name	Cluster
Carter, Cris	1	Hilton, T.Y.	2
Harrison, Marvin	1	Hopkins, DeAndre	2
Irvin, Michael	1	Humphries, Adam	2
Largent, Steve	1	Hurns, Allen	2
Maynard, Don	1	Inman, Dontrelle	2
Owens, Terrell	1	Jackson, DeSean	2
Reed, Andre	1	Jackson, Vincent	2
Rice, Jerry	1	Jeffery, Alshon	2
Taylor, Charley	1	Johnson, Steve	2
Warfield, Paul	1	Jones, Julio	2
Adams, Davante	2	Jones, Marvin	2
Aiken, Kamar	2	Kerley, Jeremy	2
Allen, Keenan	2	LaFell, Brandon	2
Amendola, Danny	2	Landry, Jarvis	2
Austin, Tavon	2	Lee, Marqise	2
Baldwin, Doug	2	Lockett, Tyler	2
Beasley, Cole	2	Maclin, Jeremy	2
Beckham, Odell	2	Marshall, Brandon	1
Benjamin, Kelvin	2	Matthews, Jordan	2
Benjamin, Travis	2	Matthews, Rishard	2
Blackmon, Justin	2	McCluster, Dexter	2
Boldin, Anquan	1	Meredith, Cameron	2
Boyd, Tyler	2	Moncrief, Donte	2
Britt, Kenny	2	Nelson, Jordy	2
Brown, Antonio	2	Parker, DeVante	2
Brown, John	2	Patterson, Cordarrelle	2
Bryant, Dez	2	Pryor, Terrelle	2
Bryant, Martavis	2	Randle, Rueben	2
Cobb, Randall	2	Roberts, Andre	2
Cooks, Brandin	2	Robinson, Allen	2
Cooper, Amari	2	Royal, Eddie	2
Crabtree, Michael	2	Sanders, Emmanuel	2
Crowder, Jamison	2	Sanu, Mohamed	2
Cruz, Victor	2	Shepard, Sterling	2
Decker, Eric	2	Shorts, Cecil	2
Diggs, Stefon	2	Smith, Steve	1
Douglas, Harry	2	Smith, Torrey	2
Edelman, Julian	2	Snead, Willie	2
Enunwa, Quincy	2	Stills, Kenny	2
Evans, Mike	2	Streater, Rod	2
Fitzgerald, Larry	1	Tate, Golden	2
Floyd, Michael	2	Thielen, Adam	2
Garcon, Pierre	2	Thomas, Demaryius	2
Ginn, Ted	2	Thomas, Michael	2
Gordon, Josh	2	Wallace, Mike	2
Green, A.J.	2	Watkins, Sammy	2
Harvin, Percy	2	Wheaton, Markus	2
Hawkins, Andrew	2	Williams, Terrance	2
Hester, Devin	2	Williams, Tyrell	2
Heyward-Bey, Darrius	2	Woods, Robert	2
Hill, Tyreek	2	Wright, Kendall	2

Our xtable shows us no difference in cluster assignment. This helps us affirm our potential predictions of Boldin, Fitzgerald, Marshall, and Smith ending up in the Hall of Fame. Next, we can move onto fuzzy clustering to see if we get any new insight from our datasets to help us make Hall of Fame predictions.

4.1.2 : Fuzzy Clustering

Our next method of clustering is a type of **soft clustering**, meaning instead of grouping data points into exactly one cluster, each element has a probability of belonging to each cluster. To find this probability, each data point has a set of membership coefficients that corresponds with the chance of being in a given cluster. This means that a data point has a probability of being in more than one cluster, which we will see through visualization. The soft method associating with this is called **fuzzy clustering**. Perhaps the most popular algorithm to use for fuzzy clustering is the **fuzzy c-means** algorithm, where the centroid point of each cluster is associated with the mean of all the data points weighed by the degree of being included in a cluster. Just like the hard clustering methods, the fuzzy c-means algorithm groups together a select amount of n data points into a C number of fuzzy clusters under some criterion. The process is very similar to k-means except now we will be considering **membership coefficients**. These values are assigned to each entry and it represents the degree of each data point belonging in each cluster. Lower coefficient values indicate a lower chance of ending up in that cluster. When we have the coefficients, we can move on and compute the fuzzy c-means algorithm, which has the following formula [6]:

$$J_m = \sum_{i=1}^{n}\sum_{j=1}^{C} u_{ij}^m ||x_i - c_j||^2 \ , \ 1 \leq m < \infty \ , \ m \in R,$$

where m is a real number greater than 1, u_{ij} represents the degree of membership to which x_i lands in cluster j, x_i is the i^{th} data point of the data, and c_j is the center point of the cluster. In order to compute u_{ij}, the membership degree, we calculate [2]

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{||x_i - c_j||}{||x_i - c_k||}\right)^{\frac{2}{m-1}}} \ .$$

The centroid point c_j is computed as [2]:

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \ x_i}{\sum_{i=1}^{N} u_{ij}^m}.$$

We repeat these two summation iterations until they converge to the maximum number of iterations and J_m converges to a local minimum. When the algorithm has converged, it means membership coefficients' change between two iterations is less than ϵ, which is the termination threshold. Just like the previous methods, this algorithm minimizes the within-cluster variances.

Working with fuzzy clustering in R, we begin with computing the membership coefficients by using the `fanny` command from the `cluster` package [13] inputting our dataset and number of desired clusters. We will begin with the quarterbacks.

Quarterbacks

With cluster 1 being the cluster that includes the Hall of Fame players, our output shows us the probability of each individual player being assigned cluster 1. Cluster 2 is excluded from the table, but we know it represents the rest of the quarterbacks. As supported by our previous clustering methods, all of the Hall of Fame players have at least a 60% chance of being placed in cluster 1 with the exception of Bob Waterfield, so we know they will always end up in that cluster. We can compute and visualize the effects of fuzzy clustering using this fuzzy c-means algorithm. To stay consistent with the other clustering methods, we will use $C = 2$ clusters and plot the clusters using `fviz_cluster` [3].

We see a lot of overlap in Figure 22, which indicates the players who have a decent probability of being within both clusters. When observing their probabilities from the previous table, we see that the majority

Name	Cluster 1 Probability	Name	Cluster 1 Probability.
Aikman, Troy	0.80	Hoyer, Brian	0.13
Dawson, Len	0.77	Johnson, Josh	0.30
Elway, John	0.78	Jones, Landry	0.15
Favre, Brett	0.67	Kaepernick, Colin	0.23
Fouts, Dan	0.81	Keenum, Case	0.11
Kelly, Jim	0.84	Luck, Andrew	0.40
Layne, Bobby	0.64	Mallett, Ryan	0.24
Marino, Dan	0.74	Manning, Eli	0.81
Namath, Joe	0.63	Manuel, EJ	0.12
Tarkenton, Fran	0.77	Mariota, Marcus	0.23
Tittle, Y.A.	0.73	McCown, Josh	0.27
Unitas, Johnny	0.79	McCown, Luke	0.18
Van Brocklin, Norm	0.63	McCoy, Colt	0.10
Warner, Kurt	0.75	Moore, Matt	0.11
Waterfield, Bob	0.40	Newton, Cam	0.48
Young, Steve	0.75	Orlovsky, Dan	0.13
Anderson, Derek	0.23	Osweiler, Brock	0.10
Bortles, Blake	0.16	Palmer, Carson	0.84
Bradford, Sam	0.38	Ponder, Christian	0.12
Brady, Tom	0.72	Prescott, Dak	0.34
Brees, Drew	0.71	Rivers, Philip	0.81
Bridgewater, Teddy	0.15	Rodgers, Aaron	0.69
Carr, Derek	0.21	Roethlisberger, Ben	0.81
Cassel, Matt	0.39	Ryan, Matt	0.79
Clemens, Kellen	0.19	Sanchez, Mark	0.31
Cousins, Kirk	0.28	Siemian, Trevor	0.14
Cutler, Jay	0.82	Smith, Alex	0.71
Dalton, Andy	0.52	Smith, Geno	0.15
Daniel, Chase	0.16	Stafford, Matthew	0.73
Fitzpatrick, Ryan	0.70	Stanton, Drew	0.22
Flacco, Joe	0.79	Tannehill, Ryan	0.38
Foles, Nick	0.17	Taylor, Tyrod	0.21
Gabbert, Blaine	0.17	Webb, Joe	0.24
Garoppolo, Jimmy	0.35	Weeden, Brandon	0.11
Glennon, Mike	0.13	Wentz, Carson	0.12
Griffin III, Robert	0.17	Wilson, Russell	0.44
Henne, Chad	0.22	Winston, Jameis	0.13
Hill, Shaun	0.13	Yates, T.J.	0.17

of the names that show up in the middle section have a variety of probability ranges. On average, these middle section players' probability to get into the first cluster and be considered for a potential Hall of Fame induction is over 20%. When looking at the clustering plot, we see the same people placed in the first cluster that we saw using the PAM technique, including Andy Dalton. Based on this, our statistical measures reaffirm the same players that have a strong chance of ending up in the Hall of Fame, with the addition of certain players that have a decent probability of also being inducted.

Running Backs

Our code will be identical to that from the quarterback section. Before visualizing the clusters, we can look at the probabilities of each player being assigned to whichever cluster using `xtable` [7].

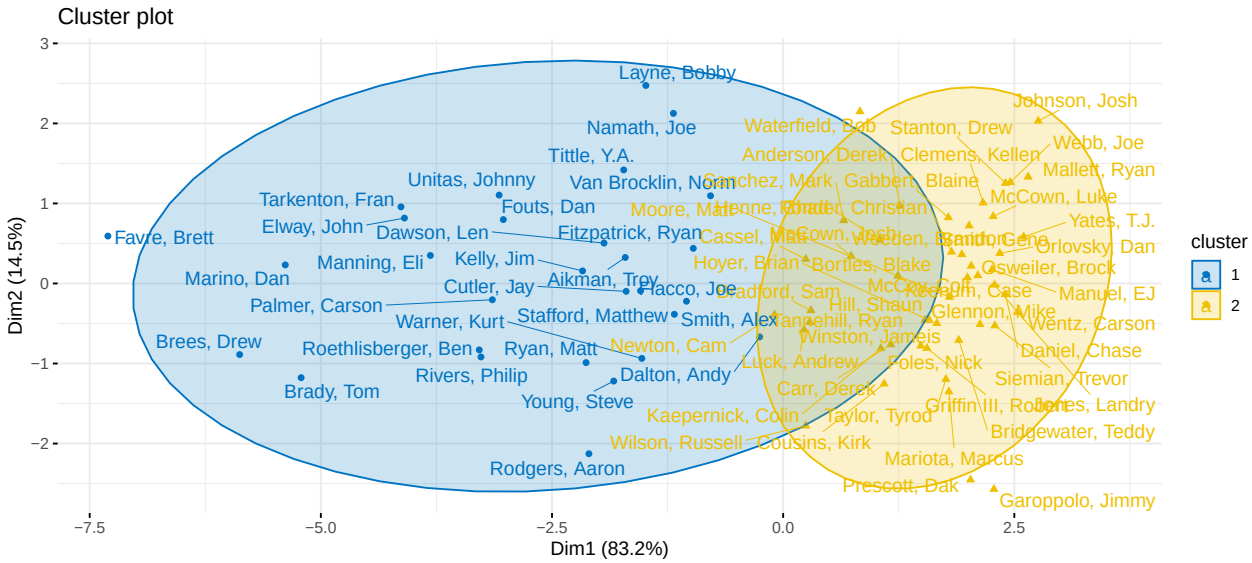

Figure 22: Fuzzy Clustering Plot for Quarterbacks

In the table, we see every player's probability of being included in cluster 1, the one with the majority of the Hall of Fame players. Apart from Floyd Little, all of the Hall of Famers have a probability of at least 67%. Unlike our previous clustering methods, we see some new players that have a pretty strong probability of ending up in cluster 1. These players include Ryan Mathews, Darren McFadden, Demarco Murray, and Jonathon Stewert. All four of these players have retired, so I will be considering each of them in my final predictions in the results section. For now, we can look at the plot produced from fuzzy clustering using `fviz_cluster` [3].

In Figure 23, we can see some of the running backs we mentioned a moment ago in the blue cluster, likely due to their relatively high probability to be included in the Hall of Fame cluster. Just like with the quarterbacks, some of the names in the overlap section have a decent probability to also be included in cluster 1. Their probabilities range from around 20% to 45%. With the addition of the new players in cluster 1, we have a bigger subset of players to consider when making my final predictions for Hall of Fame inductions.

Wide Receivers

Lastly, we have our wide receiver data. Again, the same R code will be used when computing the probabilities of each player ending up in cluster 1 and to plot the clusters themselves. We'll use the `xtable` command [7] once again

Name	Cluster 1 Probability	Name	Cluster 1 Probability
Allen, Marcus	0.70	Hyde, Carlos	0.20
Campbell, Earl	0.69	Ingram, Mark	0.42
Davis, Terrell	0.75	Ivory, Chris	0.54
Dickerson, Eric	0.71	Jennings, Rashad	0.30
Faulk, Marshall	0.72	Johnson, Chris	0.79
Harris, Franco	0.70	Johnson, David	0.25
Kelly, Leroy	0.67	Johnson, Duke	0.21
Little, Floyd	0.57	Jones, Matt	0.27
Martin, Curtis	0.71	Kelley, Rob	0.15
Payton, Walter	0.67	Lacy, Eddie	0.30
Riggins, John	0.68	Langford, Jeremy	0.21
Sanders, Barry	0.72	Lynch, Marshawn	0.79
Thomas, Thurman	0.75	Martin, Doug	0.39
Tomlinson, LaDainian	0.70	Mathews, Ryan	0.65
Abdullah, Ameer	0.15	McCoy, LeSean	0.79
Ajayi, Jay	0.34	McFadden, Darren	0.64
Allen, Javorius	0.18	McKinnon, Jerick	0.15
Anderson, C.J.	0.26	Michael, Christine	0.15
Asiata, Matt	0.28	Miller, Lamar	0.38
Bell, Joique	0.23	Morris, Alfred	0.53
Bell, Le'Veon	0.37	Murray, DeMarco	0.71
Bernard, Giovani	0.20	Murray, Latavius	0.17
Blount, LeGarrette	0.59	Oliver, Branden	0.21
Blue, Alfred	0.24	Perkins, Paul	0.15
Booker, Devontae	0.23	Peterson, Adrian	0.70
Bush, Reggie	0.62	Pierce, Bernard	0.19
Charles, Jamaal	0.62	Powell, Bilal	0.23
Coleman, Tevin	0.19	Rainey, Bobby	0.16
Crowell, Isaiah	0.23	Rawls, Thomas	0.25
Davis, Knile	0.25	Robinson, Denard	0.17
Elliott, Ezekiel	0.40	Rodgers, Jacquizz	0.19
Forsett, Justin	0.43	Sankey, Bishop	0.18
Forte, Matt	0.77	Sims, Charles	0.34
Freeman, Devonta	0.21	Spiller, C.J.	0.48
Gillislee, Mike	0.46	Starks, James	0.17
Gordon, Melvin	0.22	Stewart, Jonathan	0.73
Gore, Frank	0.69	Ware, Spencer	0.16
Gurley, Todd	0.18	West, Charcandrick	0.14
Henry, Derrick	0.21	West, Terrance	0.14
Hightower, Tim	0.35	Williams, Andre	0.29
Hill, Jeremy	0.29	Williams, DeAngelo	0.73
Hillman, Ronnie	0.13	Woodhead, Danny	0.14
Howard, Jordan	0.40	Yeldon, T.J.	0.15

The results from fuzzy clustering is vastly different from using our other clustering methods. Even though we have established Jerry Rice as arguably the best receiver to ever play, his probability to be placed in cluster 1 is the lowest out of all the Hall of Famers at just 58%. As for other players with high probabilities of cluster 1 inclusion, we see a lot more names than just the four players we've considered using the past few methods. Looking at just the players with over 50% cluster 1 probability, the names include Doug Baldwin, Odell Beckham, Antonio Brown, Dez Bryant, Michael Crabtree, Eric Decker, Pierre Garcon, A.J. Green, T.Y. Hilton, DeAndre Hopkins, DeSean Jackson, Vincent Jackson, Alshon Jeffrey, Julio Jones, Jeremy Maclin, Jordy Nelson, Emmanuel Sanders, Demaryius Thomas, and Mike Wallace. A lot of these names stick out to

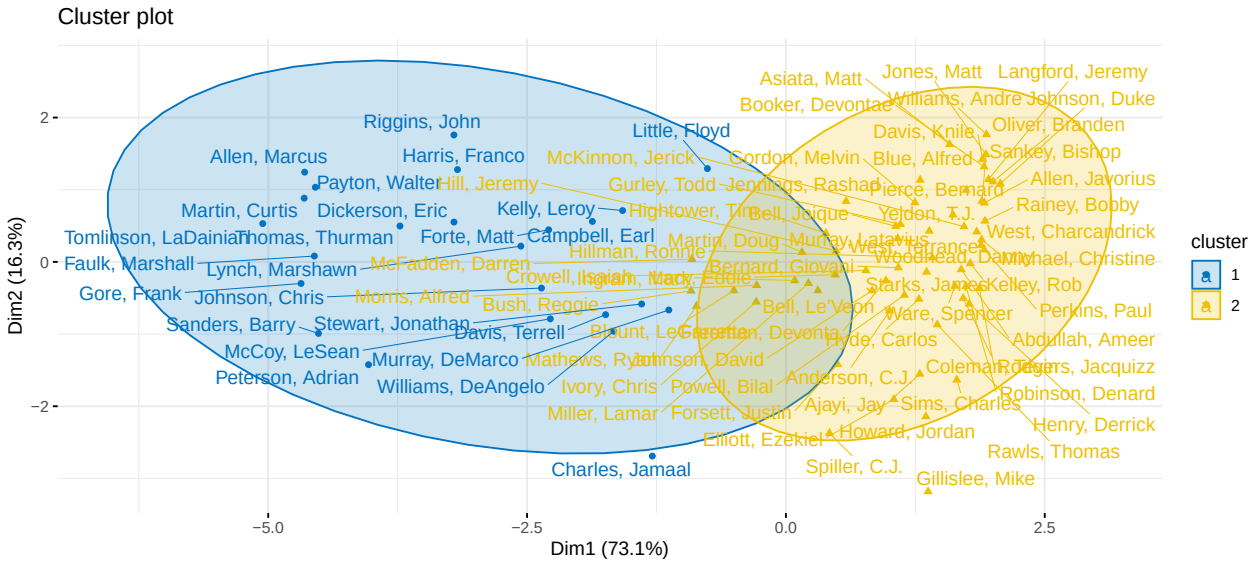

Figure 23: Fuzzy Clustering Plot for Running backs

me, but most of them have either retired or have had their careers plagued by injury.

Figure 24 is different than our previous fuzzy clustering plots such that the overlap contains a large portion of the second cluster. This leaves us with a lot of considerations to make. All of the names I just mentioned end up in the blue cluster, most of which I believe are strong candidates for Hall of Fame induction. With this many people included in the cluster, we can transition into the next section of discussing external factors not included in the dataset.

4.2 : Discussing other potential factors outside of the data

Different methods of clustering were the best indicators of pairing current NFL players with their respective position players inducted into the Hall of Fame. Based on the statistical measures taken, we have a decently sized subset of players that we can reasonably predict will be inducted into the Hall of Fame following the conclusion of their careers. However, as I briefly discussed in the previous sections, some of these predictions may not be completely accurate. For instance, if you ask anyone who has been following the NFL for years and is invested in sports and they will tell you that players like Ryan Fitzpatrick or Andy Dalton will not end up in the Hall of Fame. A potential reason for this imprecision in our clustering methods is due to factors not included within the dataset. One of the first options to consider are Super Bowl rings. From our quarterback dataset, I predict that Tom Brady will undeniably end up in the Hall of Fame due to his superior career statistics and the fact that he has 6 Super Bowl victories and other recognitions. Brady also holds multiple postseason records. Though the dataset only includes regular season statistics, Brady currently holds several records such as the all-time touchdown leader in the regular season and playoffs. He has been a true leader for the Patriots organization and is well deserving of a Hall of Fame induction. He has also won 3 regular season MVPs in his career. Out of the other quarterbacks in the cluster with Hall of Famers, Eli Manning and Ben Roethlisberger both have 2 rings and Drew Brees, Aaron Rodgers, and Joe Flacco all have 1 ring. Looking at the seasons when these quarterbacks have won it all, they played at a phenomenal level throughout the entire regular and postseason. Out of these players, only Aaron Rodgers has won the MVP award, of which he has won it twice. If they continue their success in upcoming years, it can be argued that all 5 of these quarterbacks will undoubtedly end up in the Hall of Fame.

Another interesting factor that would help contribute to making accurate Hall of Fame predictions is social media presence. Having digital data, such as total times a player's name was tweeted or maybe website polls of fans voting on the best player in the league, could have had significant effects on the clustering results. As an avid sports fan, I cannot count the number of times I have watched SportsCenter and listened to all the

Name	Cluster 1 Probability	Name	Cluster 1 Probability
Carter, Cris	0.62	Hilton, T.Y.	0.64
Harrison, Marvin	0.64	Hopkins, DeAndre	0.56
Irvin, Michael	0.67	Humphries, Adam	0.29
Largent, Steve	0.63	Hurns, Allen	0.32
Maynard, Don	0.62	Inman, Dontrelle	0.25
Owens, Terrell	0.62	Jackson, DeSean	0.63
Reed, Andre	0.65	Jackson, Vincent	0.64
Rice, Jerry	0.58	Jeffery, Alshon	0.54
Taylor, Charley	0.64	Johnson, Steve	0.46
Warfield, Paul	0.60	Jones, Julio	0.68
Adams, Davante	0.25	Jones, Marvin	0.34
Aiken, Kamar	0.21	Kerley, Jeremy	0.26
Allen, Keenan	0.30	LaFell, Brandon	0.21
Amendola, Danny	0.42	Landry, Jarvis	0.40
Austin, Tavon	0.40	Lee, Marqise	0.23
Baldwin, Doug	0.56	Lockett, Tyler	0.23
Beasley, Cole	0.31	Maclin, Jeremy	0.68
Beckham, Odell	0.54	Marshall, Brandon	0.63
Benjamin, Kelvin	0.29	Matthews, Jordan	0.35
Benjamin, Travis	0.24	Matthews, Rishard	0.26
Blackmon, Justin	0.21	McCluster, Dexter	0.39
Boldin, Anquan	0.63	Meredith, Cameron	0.26
Boyd, Tyler	0.29	Moncrief, Donte	0.28
Britt, Kenny	0.29	Nelson, Jordy	0.59
Brown, Antonio	0.68	Parker, DeVante	0.23
Brown, John	0.33	Patterson, Cordarrelle	0.39
Bryant, Dez	0.67	Pryor, Terrelle	0.22
Bryant, Martavis	0.32	Randle, Rueben	0.24
Cobb, Randall	0.43	Roberts, Andre	0.26
Cooks, Brandin	0.32	Robinson, Allen	0.35
Cooper, Amari	0.32	Royal, Eddie	0.36
Crabtree, Michael	0.60	Sanders, Emmanuel	0.54
Crowder, Jamison	0.33	Sanu, Mohamed	0.21
Cruz, Victor	0.46	Shepard, Sterling	0.33
Decker, Eric	0.56	Shorts, Cecil	0.25
Diggs, Stefon	0.24	Smith, Steve	0.65
Douglas, Harry	0.24	Smith, Torrey	0.46
Edelman, Julian	0.47	Snead, Willie	0.31
Enunwa, Quincy	0.29	Stills, Kenny	0.28
Evans, Mike	0.45	Streater, Rod	0.28
Fitzgerald, Larry	0.64	Tate, Golden	0.49
Floyd, Michael	0.40	Thielen, Adam	0.23
Garcon, Pierre	0.59	Thomas, Demaryius	0.68
Ginn, Ted	0.28	Thomas, Michael	0.23
Gordon, Josh	0.48	Wallace, Mike	0.69
Green, A.J.	0.70	Watkins, Sammy	0.41
Harvin, Percy	0.41	Wheaton, Markus	0.25
Hawkins, Andrew	0.24	Williams, Terrance	0.38
Hester, Devin	0.21	Williams, Tyrell	0.33
Heyward-Bey, Darrius	0.32	Woods, Robert	0.27
Hill, Tyreek	0.36	Wright, Kendall	0.38

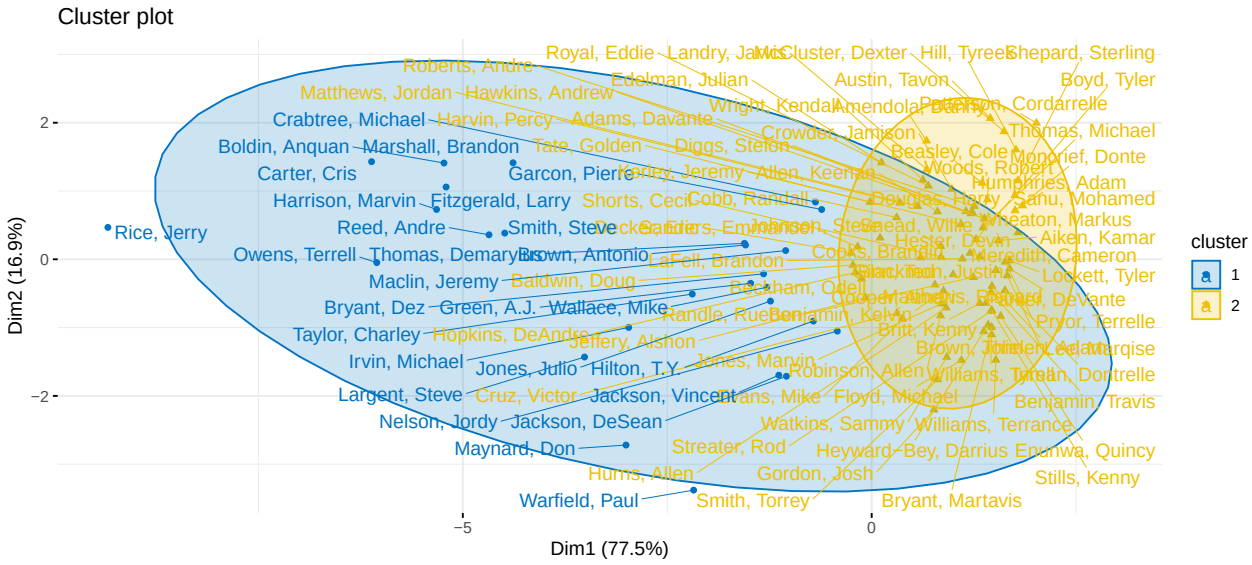

Figure 24: Fuzzy Clustering Plot for Wide Receivers

claims of Tom Brady being the GOAT, or in words, the greatest of all time. Even on Twitter, I see ESPN tweeting after every Patriots' win, and they never fail to include Brady's game stats or a picture of him. Having access to this sort of data would give me a better indication of how the rest of the world regards each player in the league. The best opinions would come from the sports analysts and those working at ESPN. These people have the best knowledge of the NFL and have been analyzing these players since their first career game.

The last factor to consider in making predictions is the occurrence of injuries or anything else that contributes to the loss of a starting job. The NFL is arguably the most dangerous professional league to play in, as players are risking their lives with every snap they play on the field. In the world of sports, there is a famous saying that goes "Next man up." To translate, when a starter gets injured, every coach has the "next man up" philosophy, meaning there is always a player to take his position. This can often lead to drastic changes for a team. Take for example a starting quarterback goes down and misses the season and the backup is called up. The former bench player has the potential of so much success that the coach may stick with that player even after the starter is healthy. This could lead to the former starting quarterback to be traded since the team found their next man up and is sticking to the player that leads them to victory. One example to consider from our dataset is Alex Smith, as he was included in the cluster with Hall of Famers. His stats included in this dataset up to 2016 has certainly put him on pace to end up among the top to ever play. However, in 2018, he suffered a gruesome compound and spiral fracture to his tibia and fibula. He has not played a game since, and now that he is 35 years old and his team has moved on with another quarterback, he is considering retirement. In other scenarios not barring injury, a starting quarterback could be performing poorly enough that he gets benched for an upcoming player. This exact situation has happened this season; Eli Manning, one of the quarterbacks who was grouped in the Hall of Fame cluster, was benched in Week 3 for rookie Daniel Jones. Andy Dalton has also been benched in favor of a rookie following an 0-8 start to the Bengals' 2019 season. This is most definitely not an uncommon occurrence, as any player's starting position can be taken at any time.

As we can see, there are several factors to take into consideration when making these types of predictions. It's definitely not easy following every player in the NFL, recording their stats, and analyzing outside factors to help make strong guesses on who will end up in the Hall of Fame. The clustering methods certainly gives us a good selection of players to chosen from. When making my final predictions, I will have to keep all these factors not represented in the dataset in mind along with the statistical measures taken.

Section 5: Results - Making final predictions

Based on our findings from using various clustering methods, along with my sports knowledge and the consideration of factors outside of the dataset, I will make final predictions on what players I believe will end up in the Hall of Fame after the conclusion of their professional careers. Each prediction will have an argument supporting my case that the player will be a future Hall of Famer. Unless indicated otherwise, all the players I have chosen were placed in the cluster that included the other Hall of Fame players.

5.1 : Quarterbacks

Tom Brady - Arguably the most logical prediction, I believe Tom Brady will undeniably end up in the Hall of Fame. The clustering methods have supported this claim, but what the dataset doesn't capture is all of his recognitions. Brady is a 14-time Pro Bowler, 3-time MVP, 6-time Super Bowl winner, and he is 2nd all-time in total passing yards and touchdown passes [1]. The sports media always claims he is the greatest quarterback of all time, and he is showing no signs of slowing down. He is 42 years old this season and has not indicated retirement anytime soon. There is no doubt Brady's name will end up in the Hall of Fame down the road.

Drew Brees - Right now, Brees is 3rd in touchdowns and 1st all time in passing yards, and these numbers are still adding [1]. He has been leading the Saints offense for over a decade, and almost every season his team has ended with a winning record. His most significant accomplishment was his Super Bowl victory in 2010. He has also rarely dealt with injuries, starting all 16 games in 12 of his seasons [1]. He missed a couple weeks this season, but is now back in the starting lineup with a winning record. Brees is another quarterback that shows very little signs of slowing down, and the Saints have always been a threat in the playoffs in seasons that he plays.

Aaron Rodgers - Considered one of the most talented quarterbacks of our generation by sports analysts everywhere, Rodgers is still making incredible throws and plays on a weekly basis at 35 years old. I have seen many tweets and SportsCenter Top 10 Plays where people make statements such as, "No other quarterback can make that type of throw." He is a 2-time MVP, a 1-time Super Bowl winner, and he is currently the all time leader in passer rating [1]. He is also in the top 10 of total touchdown passes [1]. I have seen many arguments that Rodgers is the GOAT based on his skillset. Nonetheless, there should be no concern about Aaron Rodgers not ending up in the Hall of Fame.

Eli Manning - Despite his recent struggles and being benched by a rookie this season, Eli Manning has a great resume that supports my claim. He is the reason Tom Brady doesn't have two more Super Bowl rings than he has now. He is top-10 in career passing yards and touchdowns and has two Super Bowl victories and Super Bowl MVPs under his belt [1]. The media considers him to be one of the best Giants quarterback of all time. Apart from this season, he has started every game for the Giants since 2005 and has led the team to a winning record in over half of those seasons. He is 38 years old now, but due to his longevity and success during his prime, I believe Eli will end up in the Hall of Fame.

5.2 : Running Backs

Adrian Peterson - Peterson, nicknamed AP and, was arguably the best running back in the league for several seasons during his time on the Vikings. He won the league MVP in 2012 when he rushed for over 2,000 yards, something only 6 other people have ever done in a single season [1]. He is also number 6 in all-time rushing yards and number 5 all-time in touchdowns [1]. The only downside is he has missed a lot of games in his career and suffered an ACL tear a couple years back. While he has been plagued by injuries, he is still playing today and has a chance to pass some of the greats in the Hall of Fame by advancing up the all-time lists.

Marshawn Lynch - Lynch does not have the nickname "Beast Mode" for nothing. Though he retired last year, Lynch is a Super Bowl Champion and a 5-time Pro Bowler [1]. Although he hung up the cleats at an early age of 32, he still finished his career with top-10 finishes in total yards, yards per game, and touchdowns in almost half of his seasons [1]. He also has many highlights; SportsCenter never fails to look back at some

of his best runs during the NFL season. While I wish he was still in the league, he was respected by everyone in the league, and I firmly believe he will become a future Hall of Fame inductee.

Chris Johnson - For a while, Chris Johnson had the fastest recorded 40-yard dash at the NFL combine and was regarded as the most explosive running back in the league during his prime. He won the 2009 Offensive Player of the Year when he rushed for 2,006 yards, which barely anyone has achieved as mentioned before [1]. He only ended up playing 10 seasons, but was chosen for 3 Pro Bowls and the First Team All-Pro in 2009 [1]. It's clear that he had a lot of potential, and the media would comment on how he was a once in a generation talent. I still believe that even with his small number of sucessful seasons, he will end up in the Hall of Fame.

Frank Gore - Frank Gore is 4th on the all-time rushing yard list, and the 3 players above him are all Hall of Famers [1]. Now 36 years old, Gore has played at least 10 games in every season he has played and is a 5-time Pro Bowler [1]. To put in perspective, not many running backs make it to 36 years old; I couldn't even name another one. Due to his consistency, durability, and ability as a power runner, Gore has done enough in the eyes of the media to be considered for the Hall of Fame. I believe he will end up being inducted.

5.3 : Wide Receivers

Anquan Boldin - Boldin recently retired at the age of 38, days before his 39th birthday, with a solid resume. He is a Super Bowl victor, a 3-time Pro Bowler, and was the 2003 Offensive Rookie of the Year [1]. He is top-15 in receiving yards and top-25 in receiving touchdowns [1]. He probably had one of the most consistent careers in his 14 seasons, finishing with over 1000 receiving yards in half of those seasons. He is also 9th on the all-time reception list [1]. It is only fitting that Boldin will end up in the Hall of Fame for his consistency.

Larry Fitzgerald - Fitzgerald has been in the league since 2004 and has started nearly every game in his career. He took no time to earn his status as one of the league's top receivers after recording over 1,400 receiving yards in 3 of his first 5 seasons [1]. He has also made many amazing plays during his career, as I remember always seeing him in the SportsCenter Top 10 Plays growing up. He has been chosen for the Pro Bowl 11 times and trails only Jerry Rice on the all-time receiving yardage and receptions lists [1]. He is also sixth all-time in receiving touchdowns [1]. There is no indication of when he is going to retire. He has stuck with the Arizona Cardinals for his whole career, and I believe when he finally retires as a Cardinal, his name will be called during the Hall of Fame induction ceremony in the near future.

Steve Smith - Smith played 16 NFL seasons and has earned a strong reputation throughout his career. He is a 5-time Pro Bowler and 2-time First Team All Pro [1]. What made him special was he acted as his team's punt and kick returner. He has 4 punt return touchdowns and 2 kick return touchdowns [1]. On top of that, he also is 8th all-time in receiving yards and 12th all time in receptions [1]. With his versatility, durability, and ability to make game-changing plays, Steve Smith will most likely end up in the Hall of Fame.

Deandre Hopkins - Hopkins was not initially included in the cluster with Hall of Fame players, but based on his early career, I can make a strong argument that he will end up in the Hall of Fame. He just entered the league in 2013 and has already been chosen to 3 Pro Bowls [1]. He was also placed on the All-Rookie Team in 2013 and is a 2-time First Team All Pro [1]. It is very often that I see top play replays that include Hopkins. Many NFL analysts have stated that he has the best hands in the league. Making one-hand catches is not even a surprise anymore since Hopkins has entered the league. He has already accumulated over 8,000 receiving yards in his career and has never struggled with injury. He already has multiple seasons finishing top 10 in receptions, yards, and touchdowns. If he keeps up his incredible play making, there is no doubt he will become a future Hall of Fame inductee.

Julio Jones - Jones was another receiver not initially grouped with the Hall of Fame players, but when using fuzzy clustering, the probability of being included in the Hall of Fame cluster was 68%. There is definitely a reason. The media has considered Jones to be a top-5 receiver in the league every year since I can remember. At age 30, he already has been to 6 Pro Bowls and been named to the First Team All Pro twice [1]. Out of all active players, he is top-10 in touchdowns and top-5 in receiving yards and receptions, and he finished as the yardage leader in 2015 and 2018 [1]. If he continues to be regarded as one of the league's best receivers every season, there is no doubt he will become a Hall of Fame receiver.

Works Cited

[1] "Pro Football Statistics and History." Pro-Football-Reference.Com, https://www.pro-football-reference.com/. Accessed 19 Nov. 2019.

[2] 5 Amazing Types of Clustering Methods You Should Know - Datanovia. https://www.datanovia.com/en/blog/types-of-clustering-methods-overview-and-quick-start-r-code/. Accessed 13 Nov. 2019.

[3] Alboukadel Kassambara and Fabian Mundt (2017). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.5. https://CRAN.R-project.org/package=factoextra

[4] Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

[5] Bob Rudis (2019). hrbrthemes: Additional Themes, Theme Components and Utilities for 'ggplot2'. R package version 0.6.0. https://CRAN.R-project.org/package=hrbrthemes

[6] Clustering - Fuzzy C-Means. https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html. Accessed 13 Nov. 2019.

[7] David B. Dahl, David Scott, Charles Roosen, Arni Magnusson and Jonathan Swinton (2018). xtable: Export Tables to LaTeX or HTML. R package version 1.8-3. https://CRAN.R-project.org/package=xtable

[8] Gillies, K. (2017, June). NFL Statistics, Version 1. Retrieved October 22, 2019 from ' https://www.kaggle.com/kendallgillies/nflstatistics.

[9] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

[10] Hadley Wickham (2016). gtable: Arrange 'Grobs' in Tables. R package version 0.2.0. https://CRAN.R-project.org/package=gtable

[11] Hadley Wickham, Jim Hester and Romain Francois (2017). readr: Read Rectangular Text Data. R package version 1.1.1. https://CRAN.R-project.org/package=readr

[12] K-Means Clustering Algorithm – from Wolfram MathWorld. http://mathworld.wolfram.com/K-MeansClusteringAlgorithm.html. Accessed 13 Nov. 2019.

[13] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2018). cluster: Cluster Analysis Basics and Extensions. R package version 2.0.7-1.

[14] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[15] Schubert, Erich, and Peter J. Rousseeuw. "Faster K-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms." ArXiv:1810.05691 [Cs, Stat], vol. 11807, 2019, pp. 171–87. arXiv.org, doi: 10.1007/978-3-030-32047-8_16.

[16] Simon Garnier (2018). viridis: Default Color Maps from 'matplotlib'. R package version 0.5.1. https://CRAN.R-project.org/package=viridis

[17] Taiyun Wei and Viliam Simko (2017). R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot

[18] Zilavy, Gary. "How to Calculate NFL Passer Rating Using a Formula in Excel or Google Sheets." Medium, 22 Oct. 2018, https://medium.com/@gzil/how-to-calculate-nfl-passer-rating-using-a-formula-in-excel-or-google-sheets-54eb072