Final Analysis

of

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

By

Anthony Grieco and Zachary Phillips

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

Submission Date: 05/05/2024

Foundations of Data Science and Analytics - CS345

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

Introduction

After exploring our data in our Exploratory Data Analysis, our objective shifted to building an unsupervised Machine Learning Model to answer each of our analysis questions:

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

Our goal by answering each of these questions is to identify not only why the NFL defines its current position classifications the way that it does, such as Wide Receiver or Running Back, but also how new hybrid positions within those same position groups could emerge and be formally recognized by the league in the future based solely on players’ receiving stats over the last ten years. Such hybrid positions could include currently defined sub-positions like a boundary wide receiver or a receiving back.

One thing to note is that for our model we decided to only use standardized data because both the raw and standardized sets during our Exploratory Data Analysis yielded roughly the same results regardless of the test we performed on those data sets (see line 71 of the attached R script). We also modified Taysom Hill’s position from Quarterback to Tight End because he was an outlier in our data set as our only Quarterback with at least twenty receptions (see line 12 of the attached R script). As observers of the game, we know from personal experience that Hill is most often deployed by the New Orleans Saints as a Tight-End as opposed to a Quarterback, so for the sake yielding more accurate results from our model we decided to change his position to Tight-End.

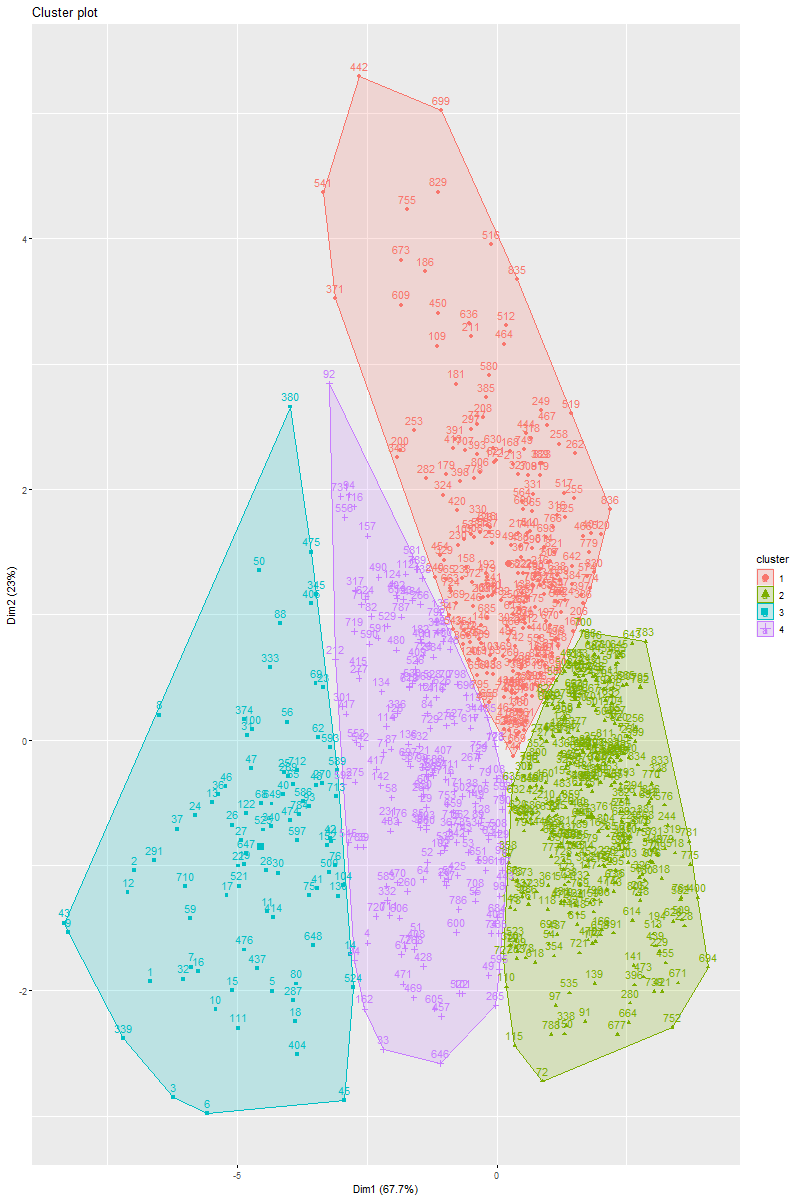
Model Development

In order to answer each of our above analysis questions, we have employed the use of multiple K-Means unsupervised machine learning models. We decided to go with an unsupervised model, and specifically K-Means, because we are really interested in predicting what future position groups in the NFL might look like and wanted a simple yet effective way to create our predicting models. For the sake of consistency, we used the same basic K-Means model to answer each of our questions but fine-tuned its parameters, such as the number of clusters it would make, based on the type of results we were looking for in that specific scenario (see example on line 107 of the attached R script). Through previous work completed in our Exploratory Data Analysis, we were able to easily identify a number of variables that we could use to guide our model including a player’s total receptions, total yards, total yards per reception, longest reception, average yards per target, average receptions per game, and average yards per game (see example on line 107 of the attached R script). These data points were derived from at most the last ten years of a player’s NFL career so that our model would better reflect how the game is currently being played in the modern era.

In each instance where we successfully used a K-Means model to cluster the data, we were able to successfully create a visualization of how our model was able to classify the data for the given scenario (see example on line 123 of attached R script). Each cluster the K-Means model was able to identify is marked by a different color, making the unsupervised model easy to understand. We were then able to interpret these visualizations to answer each of our specific analysis questions.

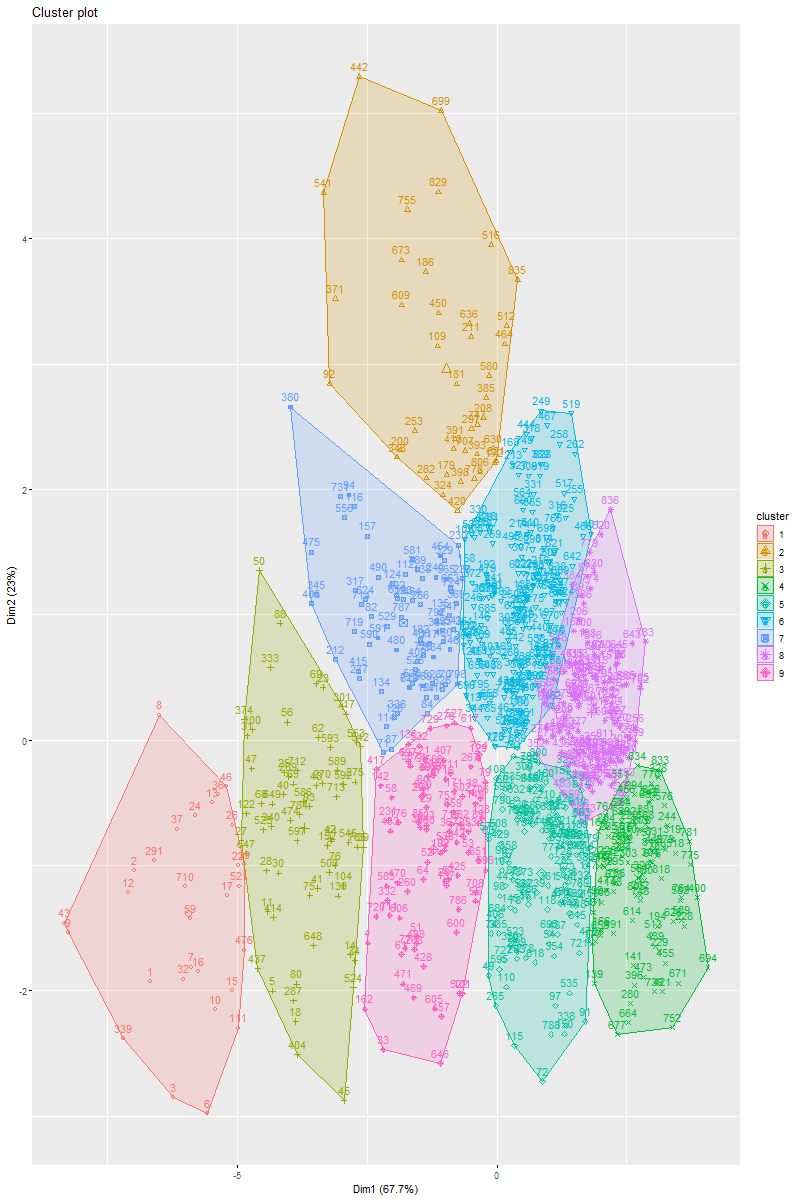
Analysis Conclusions

In reference to our first analysis question, we found that our model was able to successfully classify NFL players from the last ten years with at least twenty receptions into four distinct position group clusters (see line 121 of the attached R script). The process used to select the members of each of these clusters in our K-Means model was consistent with the variables noted above in Model Development. We decided for our model in this scenario to consist of four clusters not only because we wanted to stick with the idea that the NFL today currently has four main receiving positions groups, but also because we ran our model through a series of cluster identification tests and they suggested that four clusters was the optimal number for this scenario (see line 114 of the attached R script). Our K-Means model for our first analysis question was as follows (see line 121 of the attached R script):



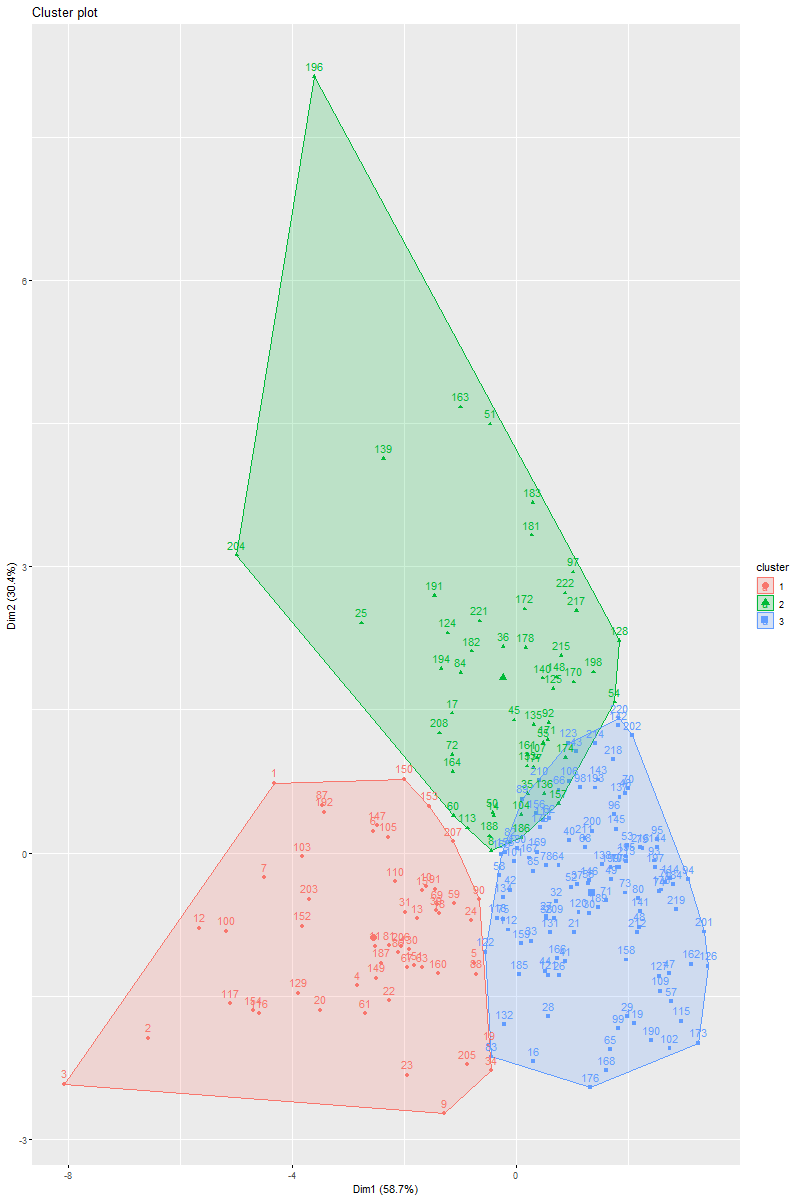
After having our model classify NFL receiver players by how similar they were to their peers regardless of their assigned position group, we compared our model’s classifications of each player to the position that they actually play in the NFL. Using a Confusion Matrix, we found that our model was only able to accurately identify a player’s actual position 35.41% of the time (see line 139 of the attached R script). This means that on 64.59% of occasions our model identified a player that plays a certain position in the NFL that doesn’t effectively match our model’s criteria for what the stats of a player who plays that position should look like in terms of a Wide Receiver, Running Back, Tight-End, or Fullback respectively. This was particularly surprising and led us to conclude that current NFL position groups are simply too broad at classifying a player’s skill set as each officially recognized position can be played in a variety of ways.

We attempted to rectify the above classification issue of current NFL players that we identified with our model by answering our second analysis question. As such, we found that our model was able to effectively classify every player with at least twenty receptions in a given year over the course of the last ten years into as many as nine separate clusters regardless of their currently assigned NFL position group. Our K-Means model for our second analysis question was as follows (see line 157 of the attached R script):



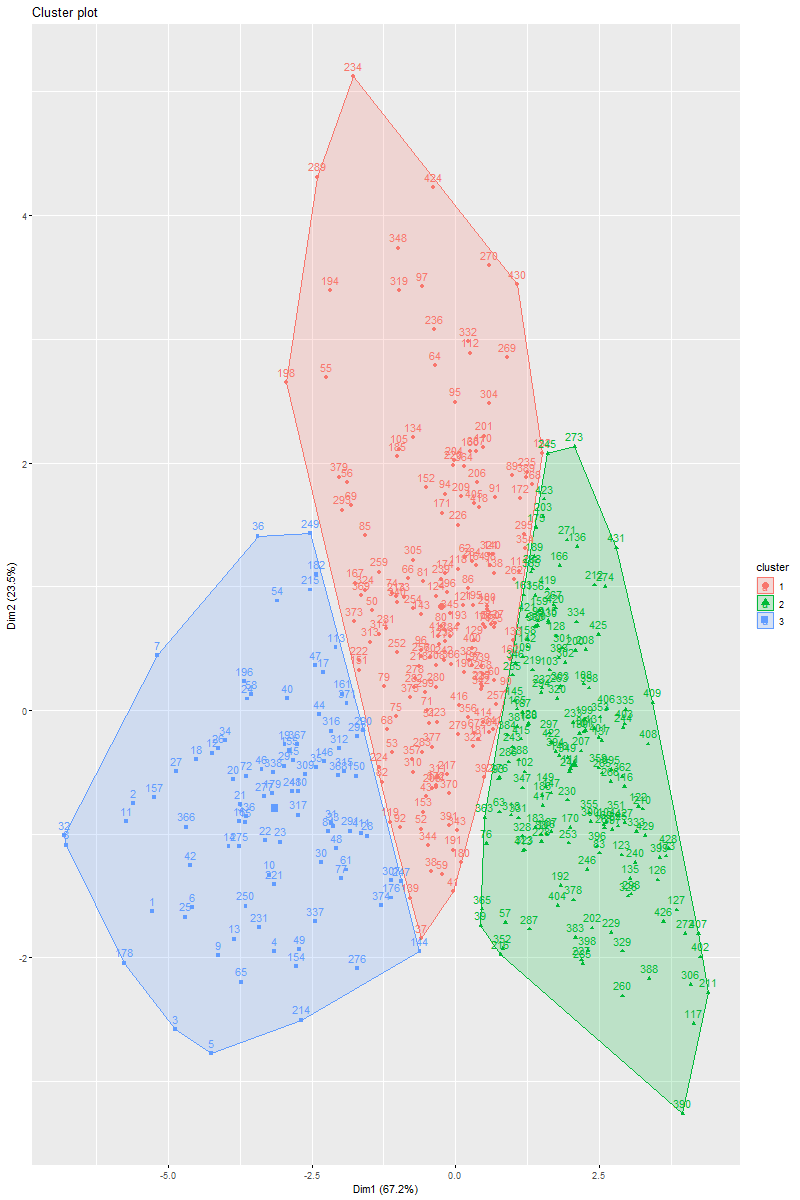
This visualization produced by our model indicates that there ought to be further classification in the NFL as to what type of Wide Receiver or Running Back a player is (see line 162 of the attached R script). For example, a player who plays more as a Hybrid or even a Receiving Running Back, like Kyren Williams of the Los Angeles Rams in Group 4 or Jahmyr Gibbs of the Detroit Lions in Group 5 respectively, are going to play the game much differently than a Running Back who qualifies as more of a Power Back like Derrick Henry formerly of the Tennessee Titans in Group 6. This model also effectively blurs the line between a player’s listed position and the picture that their performance paints of them. For example, Dalton Kincaid, officially listed as a Tight-End by the NFL, was utilized by the Buffalo Bills much more like a slot Wide Receiver as opposed to any form of a traditional Tight-End. These results, although not entirely surprising for us as scholars of the NFL, caused us to want to further explore whether clear distinctions could be made regarding sub-types of each officially recognized NFL position group.

Guided by our above findings, we sought to answer our third analysis question with the model we developed once again. Doing so allowed us to effectively find clear distinctions between each position group that are currently recognized by the NFL. For example, exclusively among players who were considered to be Running Backs we found that three distinct sub-positions could be distinguished in the data (see line 184 of the attached R script). Our K-Means model for our third analysis question when it comes to Running Backs was as follows (see line 190 of the attached R script):



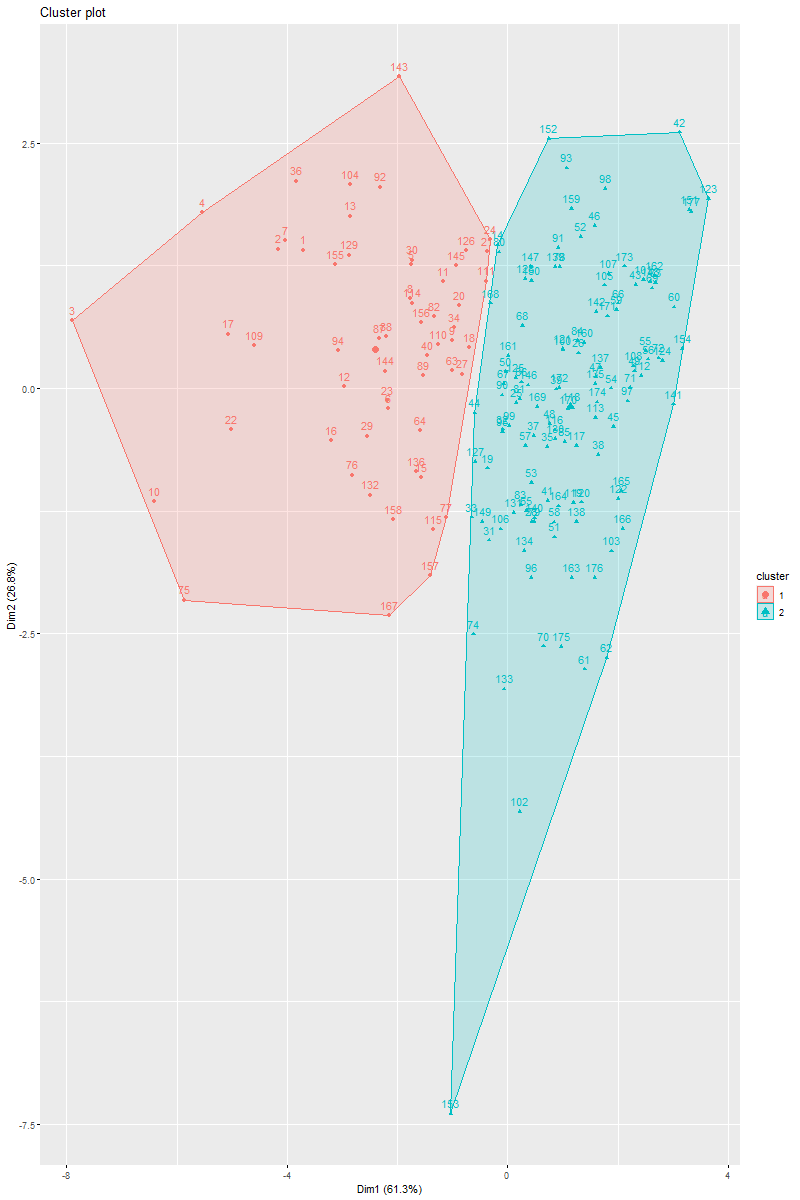
These clusters, which we have characterized to represent Receiving Backs in cluster 1, Hybrid Backs in cluster 2, and Power Backs in cluster 3 respectively (see line 195 of the attached R script), all represent different types of Running Backs in the modern NFL. Receiving Backs like Rachaad White of the Tampa Bay Buccaneers or Christian McCaffrey of the San Francisco 49ers are much more likely to be thrown the ball in short-yardage situations while also being more likely to score receiving touchdowns. Similarly, Hybrid Backs like LeSean McCoy of the Kansas City Chiefs were more likely to be thrown the ball short distances compared to traditional Running Backs but were less likely to score a touchdown on such a pass. Finally, players our model classified as Power Backs like Tyler Allgeier of the Atlanta Falcons were significantly less likely to be thrown the ball or score a touchdown on a reception than either Receiving or Hybrid Backs were.

Our model was also successfully able to identify three different types of Wide Receivers in our data set (see line 209 of the attached R script). The K-Means model we were able to produce for players classified by the league as Wide Receivers was as follows (see line 215 of the attached R script):



We have once again characterized the Wide Receiver clusters that our K-Means model produced by classifying cluster 1 as Move Receivers, cluster 2 as Slot Receivers, and cluster 3 as Boundary Receivers (see line 220 of the attached R script). An example of the type of player that our model classified as a Move Receiver would be those who fit within the mold of Stevie Johnson, who at the time was playing for the San Diego Chargers and had an above average target rate and yards per reception when compared to that of other Wide Receivers subtypes identified by our model. Slot Receivers on the other hand like Ray-Ray McCloud of the San Francisco 49ers on average had the fewest total receiving yards, touchdowns, and yards attributed to their longest reception. Finally, Boundary Receivers such as Devonta Smith of the Philadelphia Eagles generally had the greatest longest reception along with the most total yards and touchdowns out of all classically defined Wide Receivers.

We were also successful in identifying two different types of Tight-Ends with our model (see line 234 of the attached R script). The following K-Means model is what we were able to produce to distinguish different potential subtypes of Tight-Ends from one another (see line 240 of the attached R script):



Cluster 1 represents Move Tight-Ends, like Rob Gronkowski formerly of the New England Patriots, who specialize in more of a receiving role whereas cluster 2 represents Inline Tight-Ends, like Dawson Knox of the Buffalo Bills, who tend to more often be utilized as blockers in the NFL. Fittingly, Move Tight-Ends generally far outperformed Inline Tight-Ends in every receiving statistical category we looked at in our model. This distinction was particularly clear when it came to a Tight-End’s targets, receptions, total yards, receiving touchdowns, and longest reception.

The final group that we had hoped to study with our model were Fullbacks but after applying our filter ensuring that every player in our data set had to have at least twenty receptions, we realized that we only had six players who were Fullbacks that matched this criteria. This meant that we did not have enough data to run these players against our K-Means clustering model (see line 257 of the attached R script). However, this also meant that we had unintentionally created a cluster of Receiving Fullbacks from the superset of all Fullbacks that we originally had in our data set from the very beginning of our Exploratory Data Analysis. Although we were unable to clearly create a distinction between a Receiving and Blocking Fullback with our model, we were still able to clearly see that there are distinguishable types of Fullbacks beyond the NFL’s currently overarching term for the position.

Although the scope of our study was limited to identifying new sub-position groups for players who were primarily receiving players, we believe that future research that harnesses the same ideas that we encapsulated in our model could be used to help identify new sub-positions groups for every broad position currently recognized by the NFL. For example, although there are currently two main types of Cornerbacks, both Boundary and Nickel Cornerbacks are officially recognized solely as Cornerbacks in the modern NFL. An unsupervised model that is generalized for offensive players as a whole and a second model fine-tuned for defensive players could be very helpful for all 32 of the NFL’s member teams to better understand and assess the strengths and weaknesses of players who are not only on their rosters, but also prospective players to be acquired either in the NFL Draft or in Free Agency.

REFERENCES

Kassambara A, Mundt F (2020). “factoextra: Extract and Visualize the Results of Multivariate Data Analyses”. R package version 1.0.7, https://CRAN.R-project.org/package=factoextra.

Maechler et al., (2023). *cluster: Cluster Analysis Basics and Extensions*., https://CRAN.R-project.org/package=cluster.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686.

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.