Clustering Defensive Formations and Predicting Success with NFL Tracking Data

David Howell

Problem Statement:

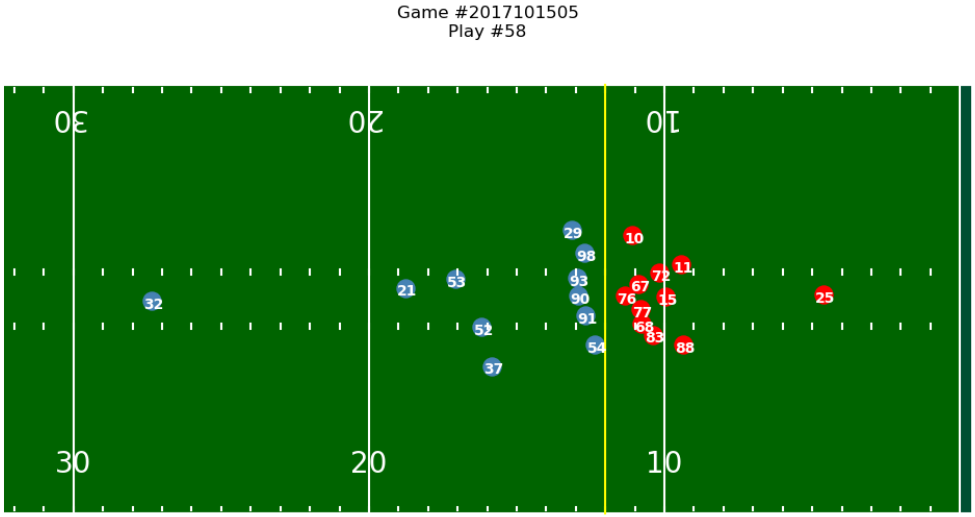
The NFL is one of the largest sports leagues in the world, with revenues reaching up to $16B/year pre-covid (Fischer, 2021). Analytics have revolutionized sports decision making in the MLB, with the rapid adoption of 'Sabermetrics', and in the NBA, with the advent of an the three-point revolution. While there has been some development in NFL strategy through analytics, in many respects, it remains behind the curve. In such a large industry, marginal gains through strategy can result in a large financial benefit.

Football is difficult to analyze. Plays outcomes are highly dependent on offensive and defensive player positioning, as well as down and distance, and field position. Moreover, the data recorded for football games has been descriptive, and of little predictive value. Football has been evolving, with passing (the more valuable offensive playcall) becoming much more prevalent than rushing, and teams choosing to use an offensive play more frequently on fourth down rather than punt, both decisions heavily favored by analytics. In recent years, the NFL has begun to track player location and movement data, and has made a portion of it available through avenues such as the Big Data Bowl, a public machine learning competition. Fans of the game have also created stats to assess the value of play outcomes. One such stat is EPA (expected points added), which assess the difference between the expected point advantage team has over their opponent before and after a play.

These relatively recent developments in football data and analysis have made analyzing the game easier and more valuable. Unsupervised learning and player location data has been used to identify pass coverages among defensive backs (Dutta et al. 2020) and assess the value of route combinations (Chu et al. 2019). The winner of the 2nd Big Data Bowl used deep learning to predict the expected number of yards gained by a rusher using the locations of players on the field at handoff (Gordeev and Singer, 2020). This model led to the NFL developing several statistics, such as Expected Rushing Yards (xRY), which are tracked by the league for every play.

Analytics and machine learning have completely transformed strategy in sports, and through this project, I aim to contribute to the advancement of football strategy through machine learning.

The first step of my project was to use clustering to identify common defensive formations. NFL defenses use several combinations of personnel and alignment to reduce predictability and counteract specific offensive playcalls. Using unsupervised learning, I grouped commonly used defensive formations. Defensive personnel is tracked in play-by-play data, but formations, which are determined by the position of personnel on the field, is not. Below is an example image generated by plotting the positions of the players on each team before the ball is snapped. The players represented by the blue dots are the defense.



The defensive formation cluster assignments were then used, along with other factors, determine the situational effectiveness of offensive plays, measured by EPA. My goal in doing this is to identify the factors that affect offensive success, and to determine if certain offensive plays perform better or worse against different defensive formations. This information can be used to assess if current NFL offensive strategy can be improved upon by choosing different playcalls or targeting certain areas of the field.

Data:

The player tracking data for this project was manually collected via a python script from the commit history of NFL Big Data Bowl. Unfortunately, the NFL deleted the player tracking data after the last iteration of the competition, which necessitated the more manual data collection approach. The data is comprised of games taking place over the first 6 weeks of the 2017 NFL season. In total, the dataset covers 90 games and over 18,540 plays. Because each player's position is tracked several times per second, there are over 30 million datapoints. Once the data is cleaned to only include defensive plays without missing data, the dataset includes over 10,000 plays.

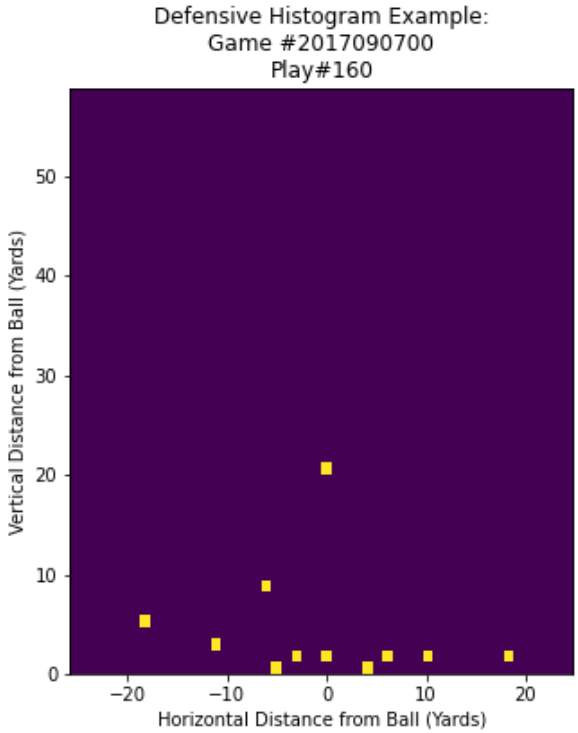
The player tracking data is augmented with data from nflfastR, an R package and data repository that collects a wide variety of stats for each NFL play. There numerous features for each play in this dataset, and these can be merged with the play data from the Big Data Bowl dataset to provide additional information, such as EPA.

Methodology:

Representation of Player Positions:

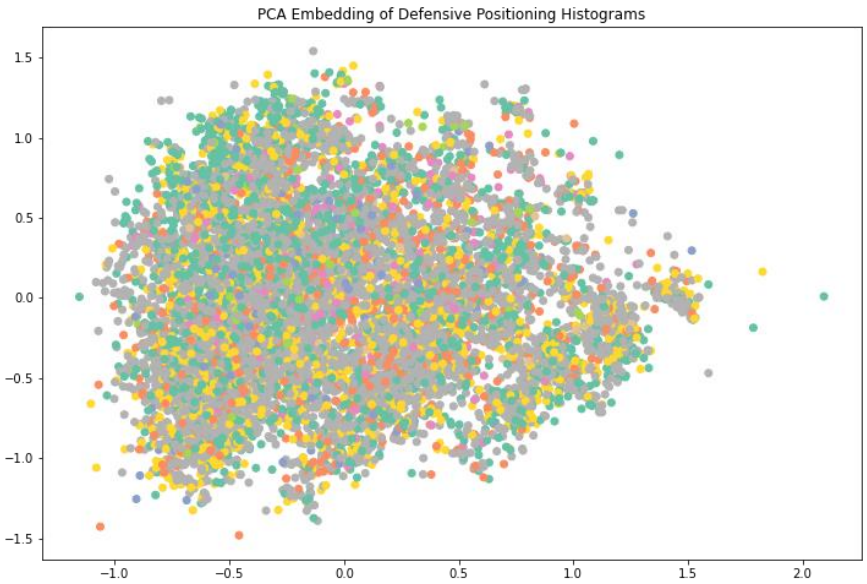
The first challenge I faced in clustering defensive formations was creating a representation of the data that could be clustered effectively. The position data for each play consists of a set of x and y coordinates for each player on the field. First, I extracted the location data for defensive players at the moment the ball was snapped in each play. The locations are given by x and y coordinates in relation to the player's position on the field, but because plays start at different positions on the field, the raw location data is not usable for determining defensive formation. Instead, I measured each player's location relative to the Center. The Center is an offensive player that snaps the ball to begin each play, so his position is a valid proxy for the location of the ball at the beginning of each play.

Because the players are represented as points on a large field, the data is very sparse. Moreover, each 'data point' is a set of x and y coordinates representing the position of the 11 defensive players. Clustering sparse, 2-dimensional datapoints proved extremely challenging. To deal with these challenges, I chose to a create a 2d histogram of each play to represent the player locations. I played around with the number of buckets along each axis, and even tried using geometric bucket sizes that increased in size as they moved farther away from the ball, to account for the increased variance in player locations farther from the line of scrimmage. In the end, I found the best results using roughly 1-yard by 1-yard buckets to create my histogram representation. The total dimension of each histogram is 50x50. An example play histogram is shown below. Each area of density shown in yellow is the location of a defensive player at the start of the play.

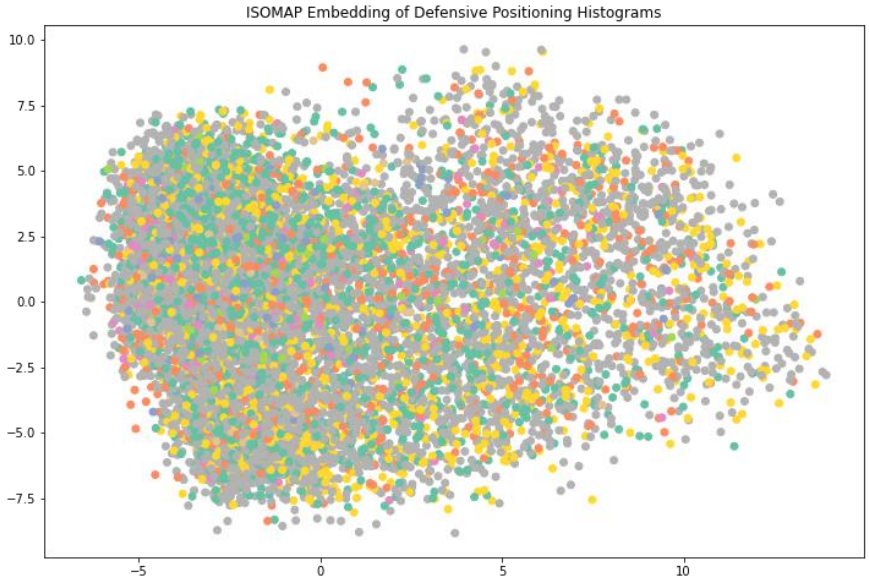


Clustering and Variable Selection:

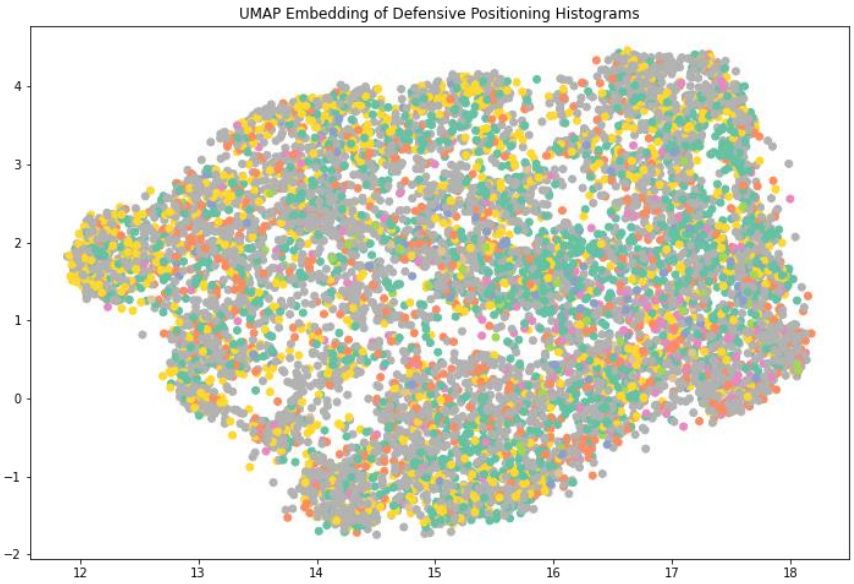
The histograms are still high-dimensional, so I used PCA, ISOMAP, and UMAP to create low-dimensional embeddings for which I could visually represent clusters. The embeddings for PCA and ISOMAP did not show any clear separations or grouping. The UMAP embedding was also fairly dense, but did show clear demarcations between groups of points. Because of its visual properties, I chose to proceed with the UMAP embedding. Plots of the embeddings are shown below. The colors correspond to defensive personnel (i.e. combinations of defensive positions). There does not appear to be much correlation between personnel and groupings on the embeddings.



The plot of the first two principal components does not show much clear separation between groups along either axis.



ISOMAP shows a dense group on the left side of the plot, and scattered points on the right, but no clear separation between groups.



UMAP shows distinct separation between clusters along the first two components. There also appears to be some evidence of grouping by color (personnel packages), though it is quite faint.

Clustering:

To cluster the data, I used both DBSCAN and K-means. DBSCAN is a density based clustering algorithm that groups together points that are closely packed together. Given the structure of the data, it appears density-based clustering is a valid approach. K-means, on the other hand, clusters by minimizing the squared Euclidean distance from each point to the nearest cluster center. Therefore, it tends to create circular groupings, and does not deal well with irregularly-shaped high-density areas. Some potential benefits of K-Means are that it assigns each point to a cluster, and it allows for a pre-defined number of clusters, which is useful if one has an underlying assumption about the number of classes in the data.

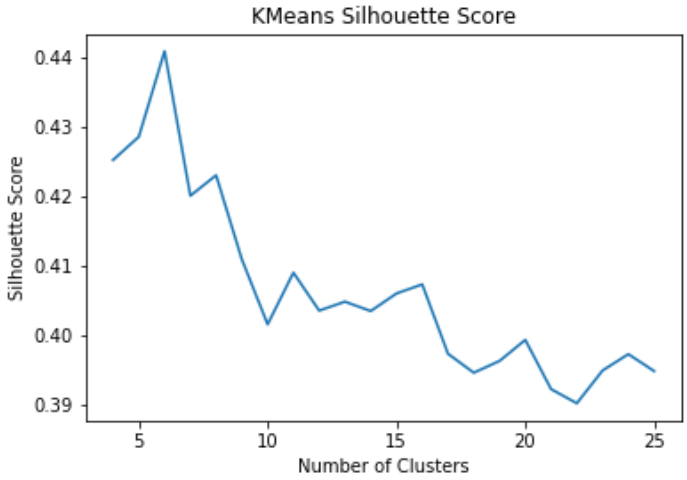
K-Means:

As mentioned above, the number of clusters is a parameter in the K-Means algorithm, which can be useful if the number of classes in the data is known. In this case, the number of distinct defensive formations is unknown. To determine the number of cluster centers, I ran the algorithm on the UMAP embedding with several different values for the n_clusters parameter. To pick the optimal value, I calculated the silhouette score of the cluster assignments for each iteration. The Silhouette Coefficient is calculated using the equation:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(b_i - a_i)}{max(a_i, b_i)}$$

Where a is the mean-intra cluster distance for each sample, b is the mean nearest-cluster distance for each sample, and n is the total number of samples.

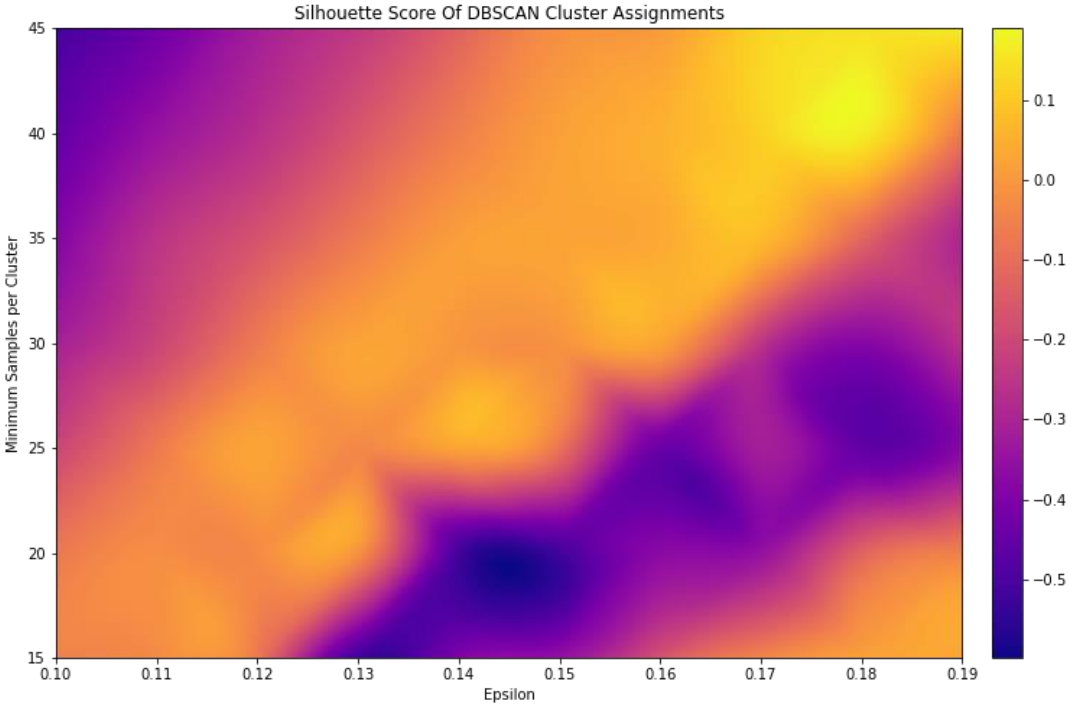
The silhouette coefficient ranges from -1 to 1, with higher scores indicating that clusters are well separated and clearly distinguished.



The highest silhouette score for K-Means was recorded when there were 6 clusters.

DBSCAN:

DBSCAN does not use a predetermined number of cluster centers, instead, it discovers clusters based on groupings of high density. The most important hyperparameters for DBSCAN are epsilon, which is the 'neighborhood' distance used to link samples together in clusters, and min_samples, which is the minimum number of samples required to define a cluster. DBSCAN does not assign all samples to a cluster. Samples that are not closely linked to nearby datapoints are considered 'noisy' samples and are not assigned. This characteristic appealed to me, as I could use the unassigned samples as the base category for my later regression. To tune the DBSCAN hyperparameters, I iterated through values of epsilon and min_samples to find the highest silhouette coefficient . A heatmap is shown below. The optimal values selected were (epsilon = 0.18, min_samples = 45).



One important consideration is that the silhouette coefficient is calculated based on the mean intra-cluster and mean nearest-cluster distances for each sample, and therefore may not be perfectly applicable to density-based clustering algorithms such as DBSCAN. I attempted to use Density Based Clustering Validation, a score that is designed for density-based clustering, but the implementation was too slow to operate on a large dataset.

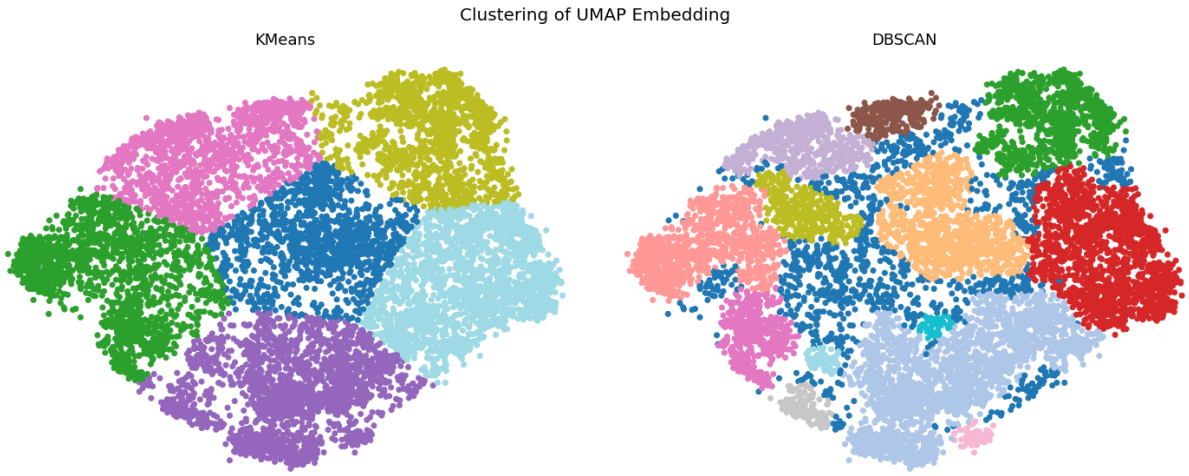
Variable Selection:

My final step was to evaluate the classifications using ANOVA, ridge regression, and forward-backward stepwise regression to determine if offensive play call success (or failure) was influenced by defensive formation clusters, and if defensive formation explains the overall success of offensive plays.

Evaluation and Final Results:

Cluster Assignments:

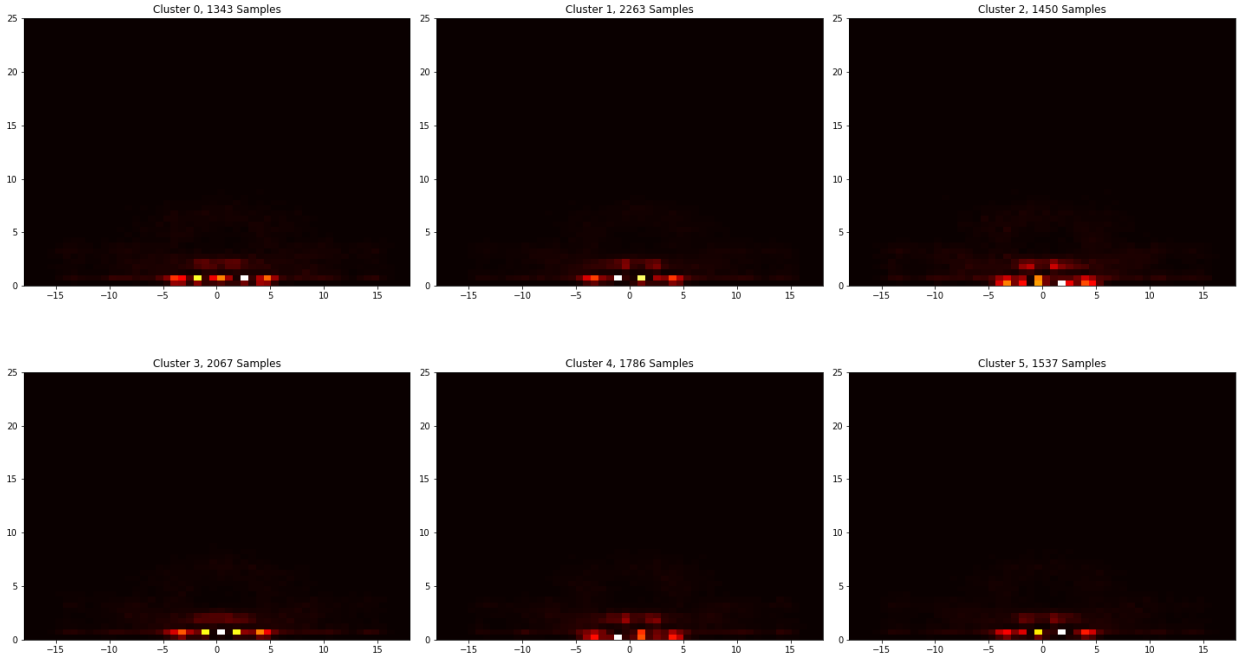
The cluster assignments from K-Means and DBSCAN are shown below. K-means was assigned 6 clusters, while DBSCAN chose 13 clusters. The noisy points that were not assigned to clusters are shown in blue.



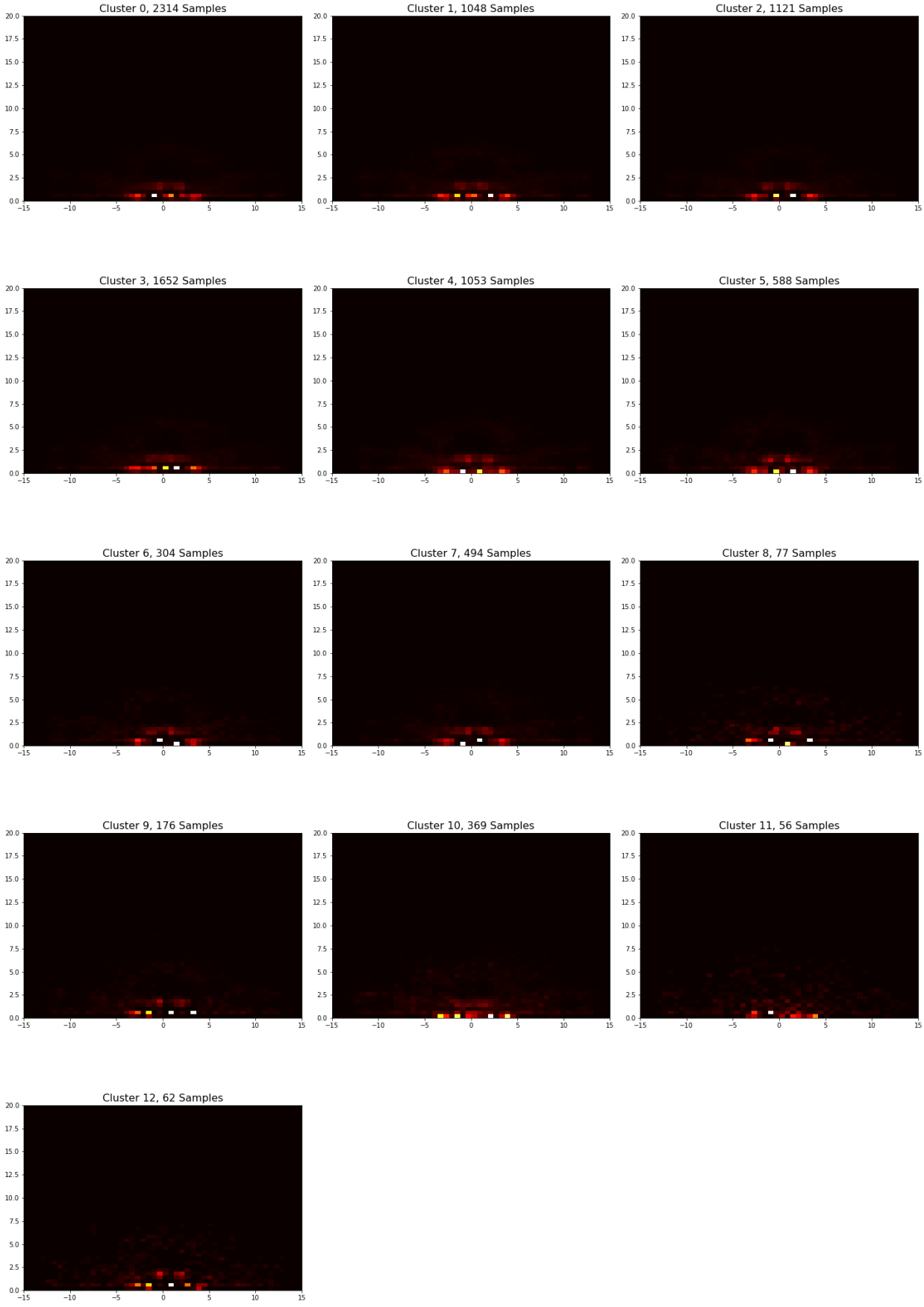
Based on the plots, DBSCAN appears to identify clusters much more effectively than K-Means. I used the cluster assignments created by each algorithm in my later analysis and found that DBSCAN assignments also had more explanatory power.

Defensive Formation Identification:

I grouped plays together by cluster assignment, and generated histograms showing the mean player positioning for each cluster. The cluster assignments for K-Means are noisier on average, which is somewhat expected, since there are fewer clusters, but also demonstrates that DBSCAN is grouping similar formations together appropriately. The mean positioning histograms do a good job capturing the positions of the 'defensive-front', players that generally start close to the line of scrimmage (defensive linemen and linebackers), but because of the variability in positioning of players who start farther from the line of scrimmage (defensive backs), the histograms fail to capture average defensive back positioning adequately.


KMeans Clustering Average Position
Cluster 0, 1343 Samples
Cluster 1, 2263 Samples
Cluster 2, 1450 Samples
Cluster 3, 2067 Samples
Cluster 4, 1786 Samples
Cluster 5, 1537 Samples

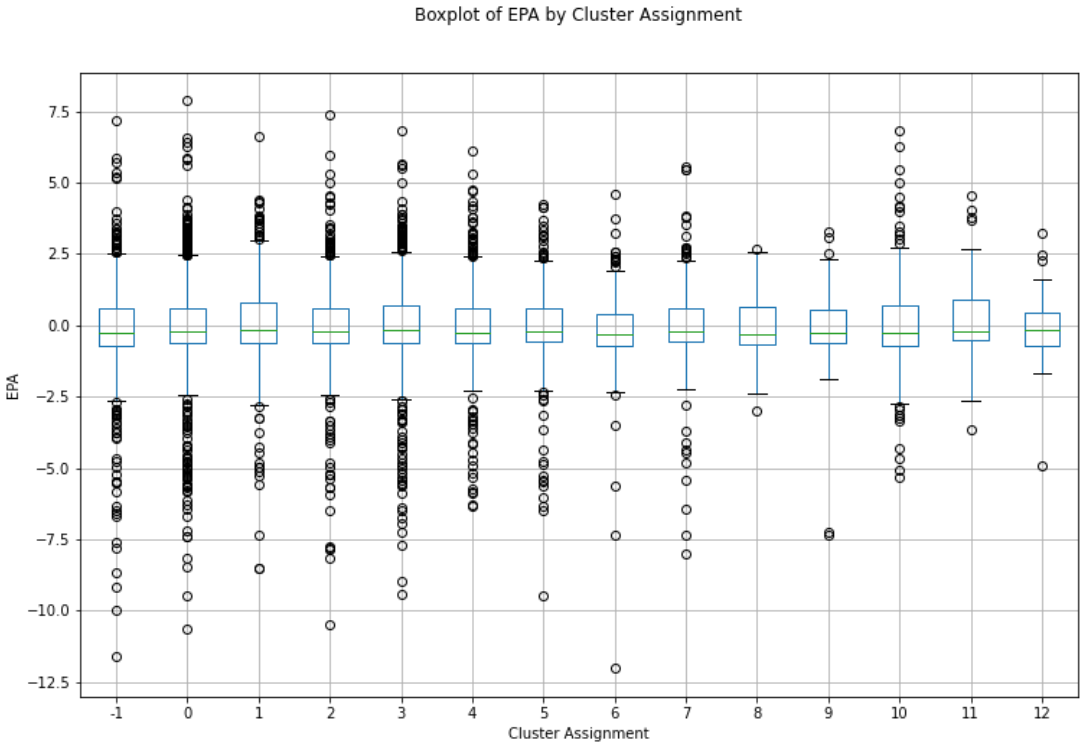
DBSCAN Clustering Average Position



Some of the clusters appear identifiable as defensive formations. For instance, DBSCAN cluster 1 shows areas of density in 5 different defensive line spots, likely corresponding to a 'Bear' front. Clusters 2 and 6 appear to show a standard 4-3 defense, with 4 defensive linemen and 3 linebackers. However, due to the variability of positioning within clusters, it is difficult to confidently equate cluster assignments to known defensive formations.

ANOVA:

To begin, my analysis, I examined if EPA/play was significantly different between different clusters. The boxplot below shows the distribution of EPA for each of the DBSCAN cluster assignments.



Keep in mind that EPA is measured from the offense's perspective. Positive values of EPA indicated that the offense had a successful result (and conversely, that the defense was unsuccessful). There does not appear to be a significant difference in distribution of EPA between the cluster assignments.

I then ran ANOVA on to determine if the mean EPA was significantly different between defensive cluster assignments. The null hypothesis for ANOVA is that the means are all the same. The alternative hypothesis is that at least two of the means are different.

```
anov1= aov(epa~cluster_, data=dat)
summary(anov1)

               Df Sum Sq Mean Sq F value Pr(>F)
cluster_       13     38   2.934   1.566 0.0868 .
Residuals   10267  19235   1.873
```

Using the p-value of 0.087, we are unable to reject the null hypothesis at even the 95% confidence level, meaning that there is not significant evidence to suggest that EPA varies based on defensive cluster assignment. I also ran a Tukey pairwise comparison to compare the differences in mean EPA between each cluster, and again found that there was not a significant difference between any of the means. The Tukey output can be found in the Appendix.

Two-Way ANOVA:

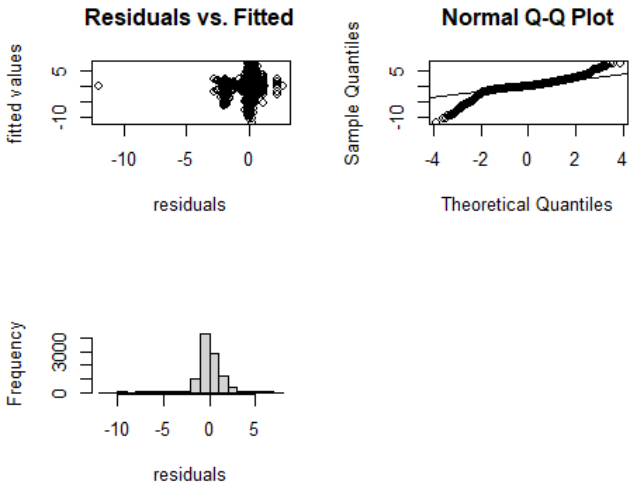
While there may not be significant variation in mean EPA between clusters, we can also evaluate if the combination of play type and cluster assignment, or the interaction between those categorical variables, has a significant effect on EPA. The two categorical variables included in the two-way ANOVA regression are cluster assignment and play type, which includes 'pass', 'run', 'run left', 'run right', 'run middle' 'pass short middle', 'pass short outside', 'pass deep middle', and 'pass deep outside'. The play classifications were created by combining play characteristics from NFLfastR.

The null hypotheses of two-way ANOVA are that the mean EPA across the levels of each of the two categorical variables are the same, and that there are no interaction effects. The alternative hypotheses are that there are differences between means across levels and there are interaction effects.

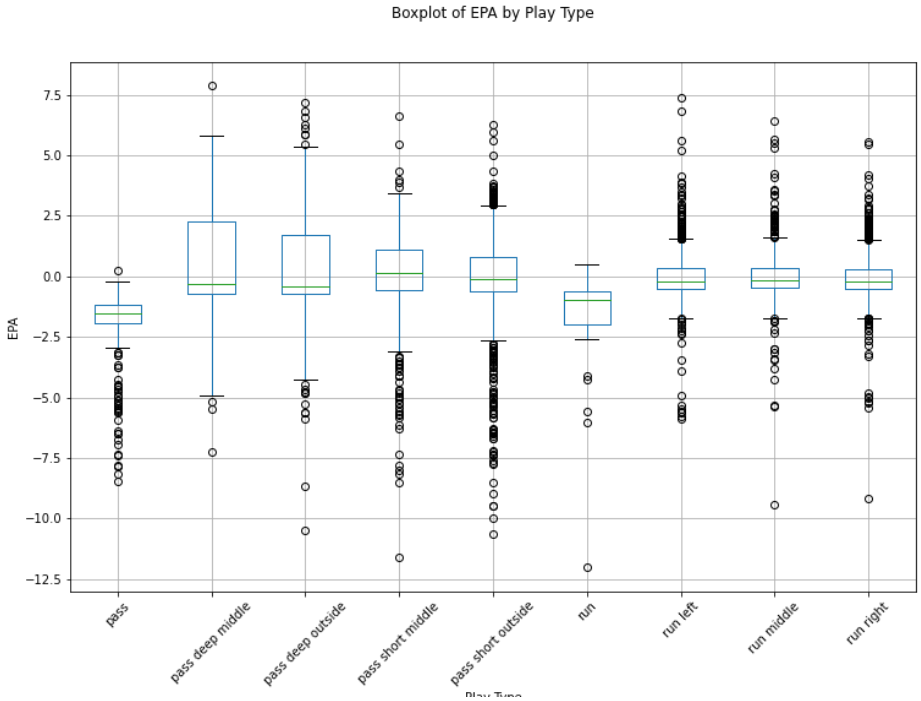
```
Two-Way ANOVA:
Observations: 10281
Dependent Variable: epa
Type: OLS linear regression

MODEL FIT:
F(121,10159) = 9.39, p = 0.00
R² = 0.10
Adj. R² = 0.09
```

The two-way ANOVA is significant, with a p-value of 0. Most of the play types were significantly different from the baseline level 'pass', with all but play type 'run', having p-values of 0. Further, several cluster-play type interactions were statistically significant with p-values of 0.01. The model output for this can be found in the appendix.

However, an examination of the model fit shows clear violations of model assumptions. ANOVA assumes that the data is independently distributed, the data is homoscedastic, and that the error terms are normally distributed.



The plot of residuals vs. fitted values shows variance increasing as fitted values increase, and clustering of fitted values, indicating that the assumptions of homoscedasticity and independence are both violated. The histogram and Q-Q plot both show severe deviations from normality. I identified that the main problem stemmed from the levels in the play type variable. Both 'pass' and 'run' had significantly lower EPAs than the other play types, as shown in the boxplot below.

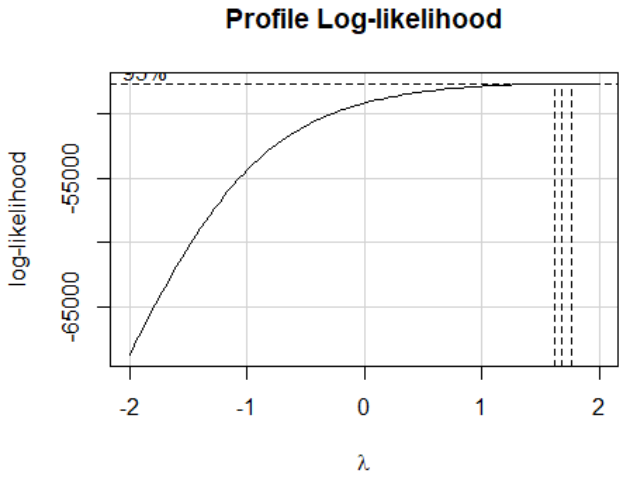


I hypothesize that the reason behind this difference is that plays that are stopped early by the defense cannot be appropriately classified by NFL recordkeepers. Therefore, a lack of information about play depth or direction is negatively correlated with play success. My goal was to determine if defensive formation affected the success of specific play calls, and so I opted to keep the more granular play information. I discarded the plays types of 'run' and 'pass' and kept those with more information. Accordingly, the remaining sample of plays is upwardly biased toward success.

I ran two-way ANOVA again on the remaining data, but the normality assumption was again violated. In an attempt to improve the distribution of errors, I used a Box-Cox transformation of EPA. Because the response has both negative and positive values, a standard Box-Cox transformation cannot be applied. To get around this, I first transformed EPA according to the following equation:

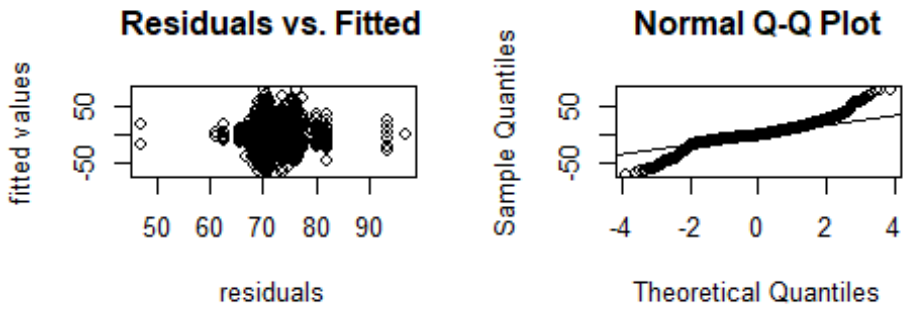
$$EPA = EPA - min(EPA) + 1$$

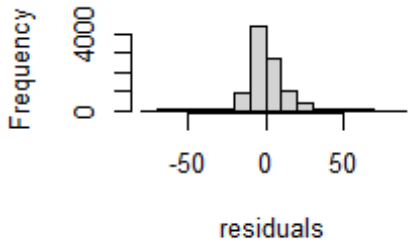
This ensured that all values of EPA were positive. I then applied a standard Box-Cox Transformation.



The optimal lambda was 1.68, which means the response in the transformed 2-way ANOVA is the positive-scaled $EPA^{1.68}$.

```
Transformed 2-Way ANOVA:
Observations: 9844
Dependent Variable: epa^1.68
Type: OLS linear regression

MODEL FIT:
F(97,9746) = 2.85, p = 0.00
R² = 0.03
Adj. R² = 0.02
```





The model assumptions appear to hold much better for this model. While errors are still not normally distributed, the fit is much better. With significant sample size, the importance of normality of errors diminishes, so I do feel comfortable using the model to help answer my question.

The overall regression is still significant, with a p-value of 0.00, indicating that we can reject at least one of the null hypotheses. Cluster 12 was statistically significantly different from the baseline level of Cluster -1 (the noisy samples that were not clustered by DBSCAN). The play types did not show statistically significant differences across levels. Several of the interaction terms were statistically significant, with p-values of 0 or below, but all they included Cluster 12, which was relatively small. It appears that Cluster 12 either displays significant explanatory power, or it included several 'big plays' with high variances in outcomes, which caused it to appear statistically significant. Given the size of the dataset and the number of levels and interaction terms, there may be statistical significance without actual real-world significance. Deeper examination of cluster 12, and the plays included in it would help answer this question. The full model output can be found in the appendix.

Variable Selection:

I used Ridge Regression and forward-backwards stepwise regression to perform variable selection and determine if cluster assignments were retained as variables in the resulting regressions. I added the following variables to control for situation, though EPA is supposed to be situationally neutral:

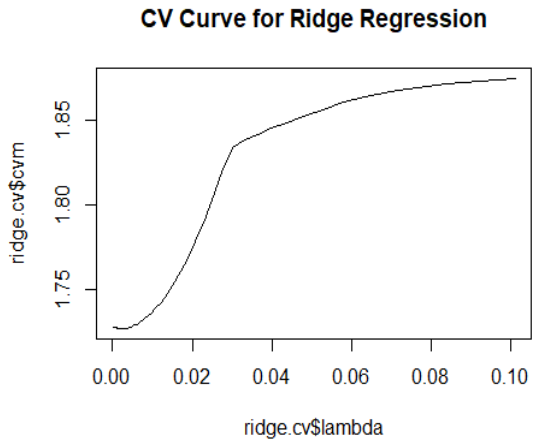
- Possession team (home or away)
- Down
- Time left in game
- Defenders in the box (a measure of players close to the line of scrimmage)

- Yards to first down
- Score differential
- Personnel

Defenders in box and personnel were included to control for closeness to the line of scrimmage and personnel, respectively, to isolate the effect of the location of players in relation to each other (defensive formation).

Ridge Regression:

I used cross-validation to determine the optimal lambda for the ridge regression.



At the optimal lambda, most of the controlling variables remained, except for possession team and several of the personnel levels. One of the play type levels was removed. Several cluster assignments were also removed. The cluster assignments do not appear to be retained in favor of any of the controlling variables. The selected variables can be found in the Appendix.

Stepwise Regression:

I performed forward-backward stepwise regression, using AIC as the complexity penalty. The variables selected were several of the play type levels, down, yards to first down, and time remaining. None of the cluster levels were selected. The model output is shown below.

```
Stepwise Model:
Observations: 9844
Dependent Variable: epa
Type: OLS linear regression

MODEL FIT:
F(9,9834) = 13.61, p = 0.00
R² = 0.01
Adj. R² = 0.01
```

```
Standard errors: OLS
------------------------------------------------------------
                              Est.   S.E.   t val.      p
--------------------------- ------- ------ -------- ------
(Intercept)                  13.30   0.11   118.33   0.00
play_type_pass deep          -0.18   0.10    -1.70   0.09
outside
play_type_pass short         -0.33   0.10    -3.23   0.00
middle
play_type_pass short         -0.48   0.10    -4.95   0.00
outside
play_type_run left           -0.57   0.10    -5.73   0.00
play_type_run middle         -0.55   0.10    -5.42   0.00
play_type_run right          -0.60   0.10    -5.96   0.00
ydstogo                      -0.02   0.00    -5.63   0.00
down                         -0.04   0.02    -2.10   0.04
game_seconds_remaining        0.00   0.00     1.55   0.12
------------------------------------------------------------
```

The overall explanatory power of the model is very low, with an $R^2$of 0.01. This suggests that the predictors available to the model, including cluster assignment, do not predict EPA well.

Conclusion, Challenges, and Further Research:

While I was able to find visually distinct clusters of player groupings, they had little to no utility in predicting EPA, and were often excluded from variable selection methods. There could be two explanations for this: either the clusters I found do not accurately represent defensive formations, or defensive formations do not significantly affect offensive success. The most difficult aspect of the project was converting player location data into clusterable representations. Because of the sparsity and multidimensionality of the data, I chose to represent each play as a histogram. I do not think that this representation resulted in true clusterings of defensive formations. I would like to revisit this project with a different methodology for reducing the data into a clusterable form, such as using an autoencoder, to see if I find better results. There are so many factors that impact play success, and the range of outcomes for a play is incredibly wide, so it was unsurprising to me that my models did not have much explanatory or predictive power. I will continue to use other methods, such as deep learning, to see if defensive formations can be accurately identified, and if play success can be predicted effectively, using player tracking data.

Data Sources:

https://github.com/nflverse/nflfastR-data/find/master

https://github.com/nfl-football-ops/Big-Data-Bowl/commits/master?before=9ad0b4b2ea36697e29e44e2399512cc1ce70358c+35&branch=master

Citations:

Beneventano, Philip, Paul D. Berger, and Bruce D. Weinberg. "Predicting run production and run prevention in baseball: the impact of Sabermetrics." *Int J Bus Humanit Technol* 2.4 (2012): 67-75.

Burke, B.. "DeepQB: Deep Learning with Player Tracking to Quantify Quarterback Decision-Making & Performance." (2019).

Chu, D., L. Wu, M. Reyers, and J. Thomson (2019): "Routes to success," NFL Big Data Bowl. https://danichusfu.github.io/files/Big_Data_Bowl.pdf.

Dutta, Rishav, Yurko, Ronald and Ventura, Samuel L. "Unsupervised methods for identifying pass coverage among defensive backs with NFL player tracking data" *Journal of Quantitative Analysis in Sports*, vol. 16, no. 2, 2020, pp. 143-161. https://doi.org/10.1515/jqas-2020-0017

Fischer, Ben. "SBJ Football: Pandemic Took 25% Bite out of NFL Revenue." Sports Business Journal, 11 Mar. 2021, www.sportsbusinessjournal.com/SB-Blogs/Newsletter-Football/2021/03/11.aspx.

Gordeev, Dmitry, and Phillip Singer. "1st Place Solution The Zoo." NFL Big Data Bowl | Kaggle, 6 Jan. 2020, www.kaggle.com/c/nfl-big-data-bowl-2020/discussion/119400.

Young, Shane. "The NBA's 3-Point Revolution Continues To Take Over." Forbes, 1 Dec. 2019, www.forbes.com/sites/shaneyoung/2019/11/30/the-nbas-three-point-revolution-continues-to-take-over.

Appendix:

Tukey Pairwise Regression Output:

```
TukeyHSD(anov1, 'cluster_', conf.level=0.99)

  Tukey multiple comparisons of means
    99% family-wise confidence level

Fit: aov(formula = epa ~ cluster_, data = dat)

$cluster_
               diff         lwr        upr     p adj
0--1   0.090475903 -0.10049436 0.28144617 0.8722620
1--1   0.196441423 -0.02956122 0.42244406 0.0556273
2--1   0.076971597 -0.14511115 0.29905435 0.9891504
3--1   0.133420545 -0.06997594 0.33681703 0.4039676
4--1   0.088442664 -0.13790374 0.31478907 0.9692354
5--1   0.054102003 -0.21391108 0.32211509 0.9999636
6--1  -0.061165639 -0.40259826 0.28026699 0.9999910
7--1   0.120875235 -0.16294351 0.40469398 0.9393950
8--1   0.099991573 -0.51985816 0.71984131 0.9999974
9--1   0.068784036 -0.35954350 0.49711157 0.9999976
10--1  0.148139576 -0.16621510 0.46249425 0.8766189
11--1  0.385551463 -0.33662885 1.10773178 0.7427717
12--1  0.023410063 -0.66409160 0.71091172 1.0000000
1-0    0.105965520 -0.09017423 0.30210527 0.7265066
2-0   -0.013504305 -0.20511421 0.17810560 1.0000000
3-0    0.042944643 -0.12665411 0.21254340 0.9995364
4-0   -0.002033239 -0.19856899 0.19450252 1.0000000
5-0   -0.036373900 -0.27973499 0.20698719 0.9999990
6-0   -0.151641542 -0.47408509 0.17080201 0.8782565
7-0    0.030399333 -0.23026652 0.29106519 1.0000000
8-0    0.009515671 -0.60008034 0.61911168 1.0000000
9-0   -0.021691867 -0.43504170 0.39165797 1.0000000
10-0   0.057663674 -0.23595593 0.35128328 0.9999735
11-0   0.295075560 -0.41832335 1.00847448 0.9515068
12-0  -0.067065840 -0.74533727 0.61120559 1.0000000
2-1   -0.119469825 -0.34601321 0.10707356 0.7589415
3-1   -0.063020877 -0.27127862 0.14523687 0.9970351
4-1   -0.107998759 -0.33872339 0.12272587 0.8819601
5-1   -0.142339420 -0.41406018 0.12938134 0.7674814
6-1   -0.257607062 -0.60195775 0.08674363 0.1965029
7-1   -0.075566188 -0.36288872 0.21175635 0.9993033
8-1   -0.096449849 -0.71791172 0.52501202 0.9999984
9-1   -0.127657387 -0.55831460 0.30299983 0.9975893
10-1  -0.048301846 -0.36582353 0.26921984 0.9999987
11-1   0.189110040 -0.53445444 0.91267453 0.9993484
12-1  -0.173031360 -0.86198687 0.51592415 0.9995757
3-2    0.056448948 -0.14754822 0.26044611 0.9988095
```

4-2	0.011471066	-0.21541527	0.23835740	1.0000000
5-2	-0.022869595	-0.29133882	0.24559963	1.0000000
6-2	-0.138137236	-0.47992803	0.20365356	0.9595673
7-2	0.043903638	-0.24034589	0.32815317	0.9999985
8-2	0.023019976	-0.59702712	0.64306708	1.0000000
9-2	-0.008187561	-0.43680066	0.42042554	1.0000000
10-2	0.071167979	-0.24357569	0.38591165	0.9998670
11-2	0.308579866	-0.41376985	1.03092959	0.9379967
12-2	-0.053561534	-0.74124114	0.63411807	1.0000000
4-3	-0.044977882	-0.25360864	0.16365287	0.9999225
5-3	-0.079318543	-0.33254790	0.17391082	0.9958333
6-3	-0.194586184	-0.52454123	0.13536886	0.5904777
7-3	-0.012545310	-0.28244747	0.25735685	1.0000000
8-3	-0.033428972	-0.64703128	0.58017333	1.0000000
9-3	-0.064636509	-0.48387221	0.35459919	0.9999985
10-3	0.014719031	-0.28713020	0.31656826	1.0000000
11-3	0.252130918	-0.46469438	0.96895621	0.9875985
12-3	-0.110010482	-0.79188483	0.57186387	0.9999974
5-4	-0.034340661	-0.30634742	0.23766609	0.9999999
6-4	-0.149608303	-0.49418471	0.19496810	0.9300875
7-4	0.032432571	-0.25516045	0.32002559	1.0000000
8-4	0.011548910	-0.61003805	0.63313587	1.0000000
9-4	-0.019658628	-0.45049635	0.41117909	1.0000000
10-4	0.059696913	-0.25806955	0.37746338	0.9999843
11-4	0.297108799	-0.42656313	1.02078073	0.9542333
12-4	-0.065032601	-0.75410095	0.62403575	1.0000000
6-5	-0.115267642	-0.48853861	0.25800333	0.9963759
7-5	0.066773232	-0.25464248	0.38818894	0.9999493
8-5	0.045889571	-0.59205139	0.68383053	1.0000000
9-5	0.014682033	-0.43943189	0.46879596	1.0000000
10-5	0.094037574	-0.25463683	0.44271198	0.9990905
11-5	0.331449460	-0.40631702	1.06921594	0.9107187
12-5	-0.030691940	-0.73454811	0.67316423	1.0000000
7-6	0.182040874	-0.20273605	0.56681780	0.8734156
8-6	0.161157212	-0.51093943	0.83325385	0.9997444
9-6	0.129949675	-0.37101408	0.63091343	0.9993988
10-6	0.209305215	-0.19851713	0.61712756	0.7924183
11-6	0.446717102	-0.32077531	1.21420952	0.6123918
12-6	0.084575702	-0.65037922	0.81953062	1.0000000
8-7	-0.020883662	-0.66562449	0.62385717	1.0000000
9-7	-0.052091199	-0.51570906	0.41152667	1.0000000
10-7	0.027264341	-0.33370091	0.38822959	1.0000000
11-7	0.264676228	-0.47897789	1.00833035	0.9862098
12-7	-0.097465172	-0.80749023	0.61255988	0.9999996
9-8	-0.031207537	-0.75134790	0.68893282	1.0000000
10-8	0.048148003	-0.61060576	0.70690177	1.0000000
11-8	0.285559890	-0.64001132	1.21113110	0.9964083
12-8	-0.076581510	-0.97535634	0.82219332	1.0000000
10-9	0.079355540	-0.40355988	0.56227096	0.9999968
11-9	0.316767427	-0.49312939	1.12666424	0.9689881

```
12-9  -0.045373973 -0.82450617 0.73375822 1.0000000
11-10  0.237411887 -0.51842357 0.99324734 0.9957180
12-10 -0.124729513 -0.84750289 0.59804387 0.9999943
12-11 -0.362141400 -1.33431777 0.61003497 0.9794293
```

2 Way ANOVA model output (untransformed):

```
MODEL FIT:
F(121,10159) = 9.39, p = 0.00
R² = 0.10
Adj. R² = 0.09

Standard errors: OLS
--------------------------------------------------------------------
                                    Est.   S.E.   t val.      p
-------------------------------- -------- ------ -------- ------
(Intercept)                        -1.81   0.17   -10.35   0.00
cluster_0                          -0.17   0.22    -0.78   0.44
cluster_1                           0.17   0.26     0.65   0.52
cluster_2                          -0.06   0.27    -0.23   0.82
cluster_3                          -0.10   0.23    -0.44   0.66
cluster_4                           0.24   0.26     0.89   0.37
cluster_5                          -0.09   0.40    -0.21   0.83
cluster_6                           0.01   0.38     0.03   0.98
cluster_7                           0.44   0.52     0.83   0.41
cluster_8                           0.08   0.68     0.12   0.90
cluster_9                           0.93   0.68     1.38   0.17
cluster_10                         -0.35   0.40    -0.87   0.38
cluster_11                          0.01   0.77     0.01   0.99
cluster_12                          0.62   0.77     0.80   0.42
play_type_pass deep middle          2.18   0.34     6.42   0.00
play_type_pass deep                 1.71   0.22     7.77   0.00
outside
play_type_pass short                1.98   0.21     9.48   0.00
middle
play_type_pass short                1.67   0.19     9.00   0.00
outside
play_type_run                       0.36   0.77     0.46   0.65
play_type_run left                  1.73   0.21     8.18   0.00
play_type_run middle                1.82   0.21     8.52   0.00
play_type_run right                 1.76   0.21     8.38   0.00
cluster_0:play_type_pass            0.24   0.42     0.57   0.57
deep middle
cluster_1:play_type_pass            0.43   0.48     0.91   0.36
deep middle
cluster_2:play_type_pass           -0.07   0.49    -0.14   0.88
deep middle
cluster_3:play_type_pass            0.08   0.44     0.19   0.85
deep middle
cluster_4:play_type_pass            0.50   0.52     0.96   0.34
```

```
deep middle
cluster_5:play_type_pass            -0.46   0.77    -0.60   0.55
deep middle
cluster_6:play_type_pass            -1.39   1.04    -1.34   0.18
deep middle
cluster_7:play_type_pass            -1.02   0.78    -1.31   0.19
deep middle
cluster_8:play_type_pass             2.22   1.50     1.48   0.14
deep middle
cluster_9:play_type_pass            -0.55   0.86    -0.64   0.52
deep middle
cluster_10:play_type_pass            0.47   0.65     0.73   0.47
deep middle
cluster_11:play_type_pass            0.80   1.55     0.52   0.60
deep middle
cluster_12:play_type_pass           -3.81   1.24    -3.07   0.00
deep middle
cluster_0:play_type_pass             0.76   0.28     2.77   0.01
deep outside
cluster_1:play_type_pass             0.22   0.33     0.68   0.50
deep outside
cluster_2:play_type_pass             0.42   0.34     1.23   0.22
deep outside
cluster_3:play_type_pass             0.42   0.29     1.43   0.15
deep outside
cluster_4:play_type_pass             0.44   0.33     1.34   0.18
deep outside
cluster_5:play_type_pass             0.35   0.47     0.74   0.46
deep outside
cluster_6:play_type_pass             0.36   0.50     0.72   0.47
deep outside
cluster_7:play_type_pass            -0.34   0.58    -0.59   0.56
deep outside
cluster_8:play_type_pass             0.62   0.87     0.71   0.48
deep outside
cluster_9:play_type_pass            -1.66   0.79    -2.09   0.04
deep outside
cluster_10:play_type_pass            1.12   0.50     2.24   0.02
deep outside
cluster_11:play_type_pass            2.39   0.93     2.57   0.01
deep outside
cluster_12:play_type_pass           -0.35   0.95    -0.37   0.71
deep outside
cluster_0:play_type_pass             0.23   0.26     0.86   0.39
short middle
cluster_1:play_type_pass            -0.04   0.31    -0.14   0.89
short middle
cluster_2:play_type_pass            -0.19   0.32    -0.60   0.55
short middle
cluster_3:play_type_pass             0.19   0.28     0.69   0.49
```

```
short middle
cluster_4:play_type_pass            -0.41   0.31    -1.31   0.19
short middle
cluster_5:play_type_pass            -0.11   0.46    -0.23   0.81
short middle
cluster_6:play_type_pass            -0.01   0.46    -0.03   0.98
short middle
cluster_7:play_type_pass            -0.90   0.57    -1.58   0.11
short middle
cluster_8:play_type_pass             0.86   0.83     1.04   0.30
short middle
cluster_9:play_type_pass            -1.06   0.72    -1.46   0.15
short middle
cluster_10:play_type_pass            0.61   0.46     1.33   0.18
short middle
cluster_11:play_type_pass           -0.30   0.90    -0.33   0.74
short middle
cluster_12:play_type_pass           -0.68   0.95    -0.72   0.47
short middle
cluster_0:play_type_pass             0.27   0.23     1.17   0.24
short outside
cluster_1:play_type_pass             0.07   0.28     0.26   0.80
short outside
cluster_2:play_type_pass             0.23   0.29     0.81   0.42
short outside
cluster_3:play_type_pass             0.32   0.25     1.28   0.20
short outside
cluster_4:play_type_pass            -0.07   0.28    -0.23   0.81
short outside
cluster_5:play_type_pass             0.26   0.42     0.63   0.53
short outside
cluster_6:play_type_pass             0.09   0.41     0.22   0.83
short outside
cluster_7:play_type_pass            -0.19   0.54    -0.35   0.73
short outside
cluster_8:play_type_pass             0.27   0.74     0.37   0.71
short outside
cluster_9:play_type_pass            -0.82   0.70    -1.18   0.24
short outside
cluster_10:play_type_pass            0.56   0.42     1.34   0.18
short outside
cluster_11:play_type_pass            0.28   0.84     0.34   0.73
short outside
cluster_12:play_type_pass           -0.49   0.84    -0.58   0.56
short outside
cluster_0:play_type_run              0.09   0.95     0.10   0.92
cluster_1:play_type_run             -1.29   1.22    -1.06   0.29
cluster_2:play_type_run              0.28   0.99     0.28   0.78
cluster_3:play_type_run              1.24   1.22     1.02   0.31
cluster_4:play_type_run              0.77   1.03     0.74   0.46
```

```
cluster_5:play_type_run              -0.71   1.08    -0.66   0.51
cluster_6:play_type_run             -10.57   1.56    -6.80   0.00
cluster_7:play_type_run              -1.49   1.30    -1.15   0.25
cluster_8:play_type_run
cluster_9:play_type_run
cluster_10:play_type_run
cluster_11:play_type_run
cluster_12:play_type_run             -0.44   1.70    -0.26   0.79
cluster_0:play_type_run               0.18   0.26     0.70   0.48
left
cluster_1:play_type_run              -0.07   0.31    -0.21   0.83
left
cluster_2:play_type_run               0.12   0.31     0.37   0.71
left
cluster_3:play_type_run               0.19   0.28     0.70   0.48
left
cluster_4:play_type_run              -0.30   0.31    -0.96   0.34
left
cluster_5:play_type_run              -0.00   0.44    -0.01   0.99
left
cluster_6:play_type_run              -0.22   0.44    -0.50   0.62
left
cluster_7:play_type_run              -0.35   0.55    -0.64   0.52
left
cluster_8:play_type_run              -0.47   0.80    -0.59   0.55
left
cluster_9:play_type_run              -0.92   0.74    -1.24   0.22
left
cluster_10:play_type_run              0.15   0.48     0.32   0.75
left
cluster_11:play_type_run             -0.03   1.52    -0.02   0.98
left
cluster_12:play_type_run             -0.65   0.87    -0.74   0.46
left
cluster_0:play_type_run               0.17   0.27     0.64   0.52
middle
cluster_1:play_type_run              -0.23   0.32    -0.73   0.46
middle
cluster_2:play_type_run              -0.06   0.32    -0.19   0.85
middle
cluster_3:play_type_run               0.13   0.28     0.46   0.65
middle
cluster_4:play_type_run              -0.36   0.31    -1.16   0.25
middle
cluster_5:play_type_run               0.20   0.45     0.46   0.65
middle
cluster_6:play_type_run              -0.14   0.45    -0.32   0.75
middle
cluster_7:play_type_run              -0.59   0.57    -1.03   0.30
middle
```

```
cluster_8:play_type_run              -0.61   0.81    -0.75   0.45
middle
cluster_9:play_type_run              -1.17   0.79    -1.47   0.14
middle
cluster_10:play_type_run             -0.05   0.48    -0.11   0.91
middle
cluster_11:play_type_run             -0.13   0.90    -0.14   0.89
middle
cluster_12:play_type_run             -0.35   1.09    -0.32   0.75
middle
cluster_0:play_type_run               0.09   0.26     0.35   0.72
right
cluster_1:play_type_run              -0.05   0.31    -0.15   0.88
right
cluster_2:play_type_run               0.08   0.31     0.26   0.80
right
cluster_3:play_type_run               0.05   0.27     0.19   0.85
right
cluster_4:play_type_run              -0.32   0.30    -1.04   0.30
right
cluster_5:play_type_run              -0.10   0.44    -0.23   0.82
right
cluster_6:play_type_run              -0.15   0.44    -0.34   0.73
right
cluster_7:play_type_run              -0.29   0.55    -0.52   0.60
right
cluster_8:play_type_run              -0.44   0.76    -0.58   0.56
right
cluster_9:play_type_run              -1.04   0.74    -1.40   0.16
right
cluster_10:play_type_run              0.29   0.46     0.64   0.52
right
cluster_11:play_type_run              0.16   0.93     0.17   0.86
right
cluster_12:play_type_run             -0.40   0.89    -0.46   0.65
right
-----------------------------------------------------------------
```

2 Way ANOVA model output (untransformed):

```
MODEL FIT:
F(97,9746) = 2.85, p = 0.00
R² = 0.03
Adj. R² = 0.02

Standard errors: OLS
-----------------------------------------------------------------
                                 Est.    S.E.   t val.      p
----------------------------- -------- ------- -------- ------
(Intercept)                     74.54    2.66    28.01   0.00
```

```
cluster_0                          1.55    3.22     0.48   0.63
cluster_1                          5.60    3.64     1.54   0.12
cluster_2                         -1.33    3.68    -0.36   0.72
cluster_3                          0.19    3.35     0.06   0.96
cluster_4                          7.55    4.07     1.86   0.06
cluster_5                         -3.56    5.95    -0.60   0.55
cluster_6                        -13.57    8.83    -1.54   0.12
cluster_7                         -5.45    5.23    -1.04   0.30
cluster_8                         22.14   12.20     1.82   0.07
cluster_9                          3.25    4.78     0.68   0.50
cluster_10                         1.96    4.61     0.43   0.67
cluster_11                         6.88   12.20     0.56   0.57
cluster_12                       -27.45    8.83    -3.11   0.00
play_type_pass deep               -4.30    2.93    -1.47   0.14
outside
play_type_pass short              -2.17    2.86    -0.76   0.45
middle
play_type_pass short              -5.29    2.72    -1.94   0.05
outside
play_type_run left                -4.93    2.88    -1.71   0.09
play_type_run middle              -4.20    2.89    -1.45   0.15
play_type_run right               -4.61    2.86    -1.61   0.11
cluster_0:play_type_pass           3.72    3.55     1.05   0.30
deep outside
cluster_1:play_type_pass          -2.44    4.06    -0.60   0.55
deep outside
cluster_2:play_type_pass           4.87    4.11     1.18   0.24
deep outside
cluster_3:play_type_pass           2.53    3.71     0.68   0.50
deep outside
cluster_4:play_type_pass          -1.42    4.45    -0.32   0.75
deep outside
cluster_5:play_type_pass           5.37    6.37     0.84   0.40
deep outside
cluster_6:play_type_pass          17.14    9.30     1.84   0.07
deep outside
cluster_7:play_type_pass           6.22    5.69     1.09   0.27
deep outside
cluster_8:play_type_pass         -15.87   13.18    -1.20   0.23
deep outside
cluster_9:play_type_pass         -10.80    6.10    -1.77   0.08
deep outside
cluster_10:play_type_pass          5.33    5.33     1.00   0.32
deep outside
cluster_11:play_type_pass         16.32   13.06     1.25   0.21
deep outside
cluster_12:play_type_pass         29.29   10.15     2.89   0.00
deep outside
cluster_0:play_type_pass          -1.32    3.48    -0.38   0.70
short middle
```

```
cluster_1:play_type_pass          -4.29    3.93    -1.09   0.27
short middle
cluster_2:play_type_pass          -1.02    3.98    -0.26   0.80
short middle
cluster_3:play_type_pass           0.39    3.62     0.11   0.91
short middle
cluster_4:play_type_pass          -9.29    4.34    -2.14   0.03
short middle
cluster_5:play_type_pass           1.54    6.29     0.24   0.81
short middle
cluster_6:play_type_pass          13.34    9.13     1.46   0.14
short middle
cluster_7:play_type_pass           1.25    5.64     0.22   0.82
short middle
cluster_8:play_type_pass         -13.50   12.94    -1.04   0.30
short middle
cluster_9:play_type_pass          -4.41    5.34    -0.83   0.41
short middle
cluster_10:play_type_pass          0.47    5.01     0.09   0.93
short middle
cluster_11:play_type_pass        -10.21   12.87    -0.79   0.43
short middle
cluster_12:play_type_pass         26.70   10.13     2.64   0.01
short middle
cluster_0:play_type_pass          -0.61    3.30    -0.18   0.85
short outside
cluster_1:play_type_pass          -3.46    3.73    -0.93   0.35
short outside
cluster_2:play_type_pass           2.81    3.77     0.75   0.46
short outside
cluster_3:play_type_pass           1.73    3.44     0.50   0.61
short outside
cluster_4:play_type_pass          -6.18    4.16    -1.48   0.14
short outside
cluster_5:play_type_pass           5.18    6.04     0.86   0.39
short outside
cluster_6:play_type_pass          14.23    8.93     1.59   0.11
short outside
cluster_7:play_type_pass           7.64    5.36     1.43   0.15
short outside
cluster_8:play_type_pass         -19.20   12.48    -1.54   0.12
short outside
cluster_9:play_type_pass          -2.28    5.03    -0.45   0.65
short outside
cluster_10:play_type_pass         -0.08    4.74    -0.02   0.99
short outside
cluster_11:play_type_pass         -4.41   12.53    -0.35   0.72
short outside
cluster_12:play_type_pass         28.44    9.28     3.07   0.00
short outside
```

```
cluster_0:play_type_run          -1.53    3.46    -0.44   0.66
left
cluster_1:play_type_run          -4.69    3.93    -1.19   0.23
left
cluster_2:play_type_run           1.75    3.93     0.45   0.66
left
cluster_3:play_type_run           0.63    3.60     0.18   0.86
left
cluster_4:play_type_run          -8.14    4.33    -1.88   0.06
left
cluster_5:play_type_run           2.73    6.15     0.44   0.66
left
cluster_6:play_type_run          11.43    9.05     1.26   0.21
left
cluster_7:play_type_run           6.11    5.46     1.12   0.26
left
cluster_8:play_type_run         -26.00   12.81    -2.03   0.04
left
cluster_9:play_type_run          -3.32    5.55    -0.60   0.55
left
cluster_10:play_type_run         -3.96    5.16    -0.77   0.44
left
cluster_11:play_type_run         -7.40   17.08    -0.43   0.66
left
cluster_12:play_type_run         27.00    9.53     2.83   0.00
left
cluster_0:play_type_run          -1.60    3.49    -0.46   0.65
middle
cluster_1:play_type_run          -6.27    3.99    -1.57   0.12
middle
cluster_2:play_type_run           0.23    3.99     0.06   0.95
middle
cluster_3:play_type_run           0.26    3.65     0.07   0.94
middle
cluster_4:play_type_run          -8.70    4.33    -2.01   0.04
middle
cluster_5:play_type_run           4.75    6.20     0.77   0.44
middle
cluster_6:play_type_run          12.22    9.11     1.34   0.18
middle
cluster_7:play_type_run           4.09    5.58     0.73   0.46
middle
cluster_8:play_type_run         -26.96   12.87    -2.09   0.04
middle
cluster_9:play_type_run          -5.51    6.08    -0.91   0.37
middle
cluster_10:play_type_run         -5.60    5.22    -1.07   0.28
middle
cluster_11:play_type_run         -8.09   12.87    -0.63   0.53
middle
```

```
cluster_12:play_type_run         29.94   11.24     2.66   0.01
middle
cluster_0:play_type_run          -2.50    3.45    -0.73   0.47
right
cluster_1:play_type_run          -4.65    3.94    -1.18   0.24
right
cluster_2:play_type_run           1.28    3.93     0.33   0.75
right
cluster_3:play_type_run          -0.84    3.60    -0.23   0.82
right
cluster_4:play_type_run          -8.45    4.29    -1.97   0.05
right
cluster_5:play_type_run           1.64    6.17     0.27   0.79
right
cluster_6:play_type_run          12.00    9.05     1.33   0.19
right
cluster_7:play_type_run           6.73    5.48     1.23   0.22
right
cluster_8:play_type_run         -25.77   12.58    -2.05   0.04
right
cluster_9:play_type_run          -4.51    5.54    -0.81   0.42
right
cluster_10:play_type_run         -2.73    5.03    -0.54   0.59
right
cluster_11:play_type_run         -4.85   13.04    -0.37   0.71
right
cluster_12:play_type_run         29.15    9.65     3.02   0.00
right
-------------------------------------------------------------
```

Ridge Regression Output:

```
ridge.model <- glmnet(X, data[,10], alpha=1, nlambda = 100)
coef(ridge.model, ridge.cv$lambda.min)

48 x 1 sparse Matrix of class "dgCMatrix"
                                      s1
(Intercept)                1.292226e+01
cluster_-1                -6.623211e-02
cluster_0                  .
cluster_1                  9.055690e-02
cluster_2                  .
cluster_3                  3.568258e-02
cluster_4                  .
cluster_5                 -2.109475e-02
cluster_6                 -5.993729e-02
cluster_7                  .
cluster_8                  .
cluster_9                 -4.323430e-02
cluster_10                 3.325088e-02
cluster_11                 2.238151e-01
```

```
cluster_12                    .
play_type_pass deep outside   1.375942e-01
play_type_pass short middle   .
play_type_pass short outside -1.355191e-01
play_type_run left           -2.171859e-01
play_type_run middle         -1.958001e-01
play_type_run right          -2.397033e-01
posteam_typehome              .
game_seconds_remaining        1.461961e-05
defendersInTheBox            -1.297120e-03
down                         -2.309121e-02
ydstogo                      -1.615554e-02
score_differential            5.280475e-04
Personnel227                  .
Personnel245                 -2.675081e-01
Personnel308                  .
Personnel317                 -2.394105e-01
Personnel326                  5.019641e-03
Personnel335                  3.082774e-02
Personnel416                  .
Personnel425                  .
Personnel434                 -2.729630e-02
Personnel443                 -3.463810e-01
Personnel515                  1.762641e-01
Personnel524                  .
Personnel533                  .
Personnel542                  2.415587e-01
Personnel614                  .
Personnel623                  9.724342e-02
Personnel632                  1.059421e-01
Personnel641                  1.925971e-01
Personnel713                 -2.708879e+00
Personnel722                  .
Personnel731                  .
```

Stepwise Regression output:

```
MODEL INFO:
Observations: 9844
Dependent Variable: epa
Type: OLS linear regression

MODEL FIT:
F(9,9834) = 13.61, p = 0.00
R² = 0.01
Adj. R² = 0.01

Standard errors: OLS
-------------------------------------------------------------
                               Est.   S.E.   t val.      p
```

```
--------------------------- ------- ------ -------- ------
(Intercept)                  13.30   0.11   118.33   0.00
play_type_pass deep          -0.18   0.10    -1.70   0.09
outside
play_type_pass short         -0.33   0.10    -3.23   0.00
middle
play_type_pass short         -0.48   0.10    -4.95   0.00
outside
play_type_run left           -0.57   0.10    -5.73   0.00
play_type_run middle         -0.55   0.10    -5.42   0.00
play_type_run right          -0.60   0.10    -5.96   0.00
ydstogo                      -0.02   0.00    -5.63   0.00
down                         -0.04   0.02    -2.10   0.04
game_seconds_remaining        0.00   0.00     1.55   0.12
-----------------------------------------------------------

step(base_mod, scope=list(lower=base_mod, upper=full_mod), direction="both",
k=2)

Start:  AIC=5356.19
epa ~ 1

                        Df Sum of Sq   RSS    AIC
+ play_type_             6   150.190 16808 5280.6
+ ydstogo                1    32.635 16926 5339.2
<none>                               16958 5356.2
+ down                   1     2.373 16956 5356.8
+ game_seconds_remaining 1     2.199 16956 5356.9
+ defendersInTheBox      1     2.082 16956 5357.0
+ score_differential     1     0.051 16958 5358.2
+ posteam_type           1     0.003 16958 5358.2
+ cluster_              13    36.227 16922 5361.1
+ Personnel             21    61.499 16897 5362.4

Step:  AIC=5280.62
epa ~ play_type_

                        Df Sum of Sq   RSS    AIC
+ ydstogo                1    46.555 16762 5255.3
+ game_seconds_remaining 1     4.390 16804 5280.0
<none>                               16808 5280.6
+ defendersInTheBox      1     3.339 16805 5280.7
+ score_differential     1     1.176 16807 5281.9
+ down                   1     0.125 16808 5282.5
+ posteam_type           1     0.007 16808 5282.6
+ Personnel             21    61.079 16747 5286.8
+ cluster_              13    33.177 16775 5287.2
- play_type_             6   150.190 16958 5356.2

Step:  AIC=5255.31
```

```
epa ~ play_type_ + ydstogo

                         Df Sum of Sq   RSS    AIC
+ down                    1     7.812 16754 5252.7
+ game_seconds_remaining  1     4.418 16757 5254.7
<none>                                16762 5255.3
+ score_differential      1     1.499 16760 5256.4
+ defendersInTheBox       1     0.090 16762 5257.3
+ posteam_type            1     0.045 16762 5257.3
+ cluster_               13    32.249 16729 5262.4
+ Personnel              21    53.901 16708 5265.6
- ydstogo                 1    46.555 16808 5280.6
- play_type_              6   164.110 16926 5339.2

Step:  AIC=5252.72
epa ~ play_type_ + ydstogo + down

                         Df Sum of Sq   RSS    AIC
+ game_seconds_remaining  1     4.100 16750 5252.3
<none>                                16754 5252.7
+ score_differential      1     1.763 16752 5253.7
+ defendersInTheBox       1     0.144 16754 5254.6
+ posteam_type            1     0.055 16754 5254.7
- down                    1     7.812 16762 5255.3
+ cluster_               13    33.513 16720 5259.0
+ Personnel              21    53.144 16701 5263.4
- ydstogo                 1    54.242 16808 5282.5
- play_type_              6   171.864 16926 5341.2

Step:  AIC=5252.31
epa ~ play_type_ + ydstogo + down + game_seconds_remaining

                         Df Sum of Sq   RSS    AIC
<none>                                16750 5252.3
- game_seconds_remaining  1     4.100 16754 5252.7
+ score_differential      1     1.460 16748 5253.5
+ defendersInTheBox       1     0.259 16750 5254.2
+ posteam_type            1     0.084 16750 5254.3
- down                    1     7.493 16757 5254.7
+ cluster_               13    34.032 16716 5258.3
+ Personnel              21    53.576 16696 5262.8
- ydstogo                 1    53.988 16804 5282.0
- play_type_              6   173.832 16924 5342.0


Call:
lm(formula = epa ~ play_type_ + ydstogo + down + game_seconds_remaining,
    data = data)

Coefficients:
```

```
                (Intercept)   play_type_pass deep outside
                  1.330e+01                    -1.777e-01
play_type_pass short middle  play_type_pass short outside
                  -3.283e-01                   -4.776e-01
         play_type_run left          play_type_run middle
                  -5.732e-01                   -5.525e-01
        play_type_run right                      ydstogo
                  -5.966e-01                   -1.979e-02
                        down        game_seconds_remaining
                  -3.741e-02                    1.958e-05
```