Finding Unique Plays Using Clustering Algorithms

Joseph Harter

University of Colorado Boulder

DTSA-5011 Summer 1 2025

June 19, 2025

Repo: https://www.github.com/joe-harter/dtsa-5510-final

Abstract

In this paper I'll explore how to use unsupervised machine learning algorithms to detect patterns in the formations of NFL[1] offenses. The data is from the first week of tracking data for the 2022 NFL season. Being able to reduce the number of formations when analyzing thousands of NFL plays can help focus analysis and speed up other models. I first identify features of an NFL formation (number of players, and x,y location of those players). Then I feed that into several different clustering algorithms and use relevant scores to determine which algorithm is the most useful. AgglomerativeClustering and KMeans produced identical clusters which makes sense given the features chosen. I identified 50 unique clusters with the given dataset. This method identified 14 outlier formations which suggests clustering methods may be a useful way to identify innovative plays/formations in NFL games. The dataset is part of a completed kaggle competition but can still be accessed.[2]

Keywords: clustering, NFL, kaggle

[1] Wikipedia contributors. (2025, June 13). *National Football League*. Wikipedia. https://en.wikipedia.org/wiki/National_Football_League
[2] *NFL Big Data Bowl 2025*. (n.d.). Kaggle. https://www.kaggle.com/competitions/nfl-big-data-bowl-2025/data

Finding Unique Plays Using Clustering Algorithms

I'll walk through highlights of the exploration, cleaning, and results from two different selections of features trained on multiple clustering algorithms. This paper gives a good "no code" summary of the work I did. But the related code can be found in the repository mentioned above. The files "tracking_eda.ipynb" and "plays_eda.ipynb" hosts all exploratory analysis. "Model.ipynb" and "model_smaller_features.ipynb" hosts all code for the unsupervised models as well as the different preprocessing steps for each.

Brief Primer On Offensive Formations

This paper assumes some familiarity with American football. Most of this paper really only deals with where players are located on the field of play though so the following chart may be informative. For more in depth rules the NFL's Rookie Guide may be informative.[3]

The offense is the team that is trying to score points by moving towards the opponents end of the field to either get the ball into the endzone or kick the ball through the yellow uprights. There are 11 players. The center is the only player who can toss (snap) the football back to the quarterback to start the play. There must be at least 6 other players along the same lateral line as the center when the ball is snapped. The quarterback can then give the ball (hand off) to another player, run it himself, or throw it to an eligible player (receiver). These players tend to be non-blocking positions like tight-end, running back, or wide receiver. See figure 1 for a common formation and take note of the colors. These colors will be used in further visualizations.

---

[3] *Rookie's Guide | NFL Football Operations*. (n.d.). https://operations.nfl.com/learn-the-game/nfl-basics/rookies-guide/

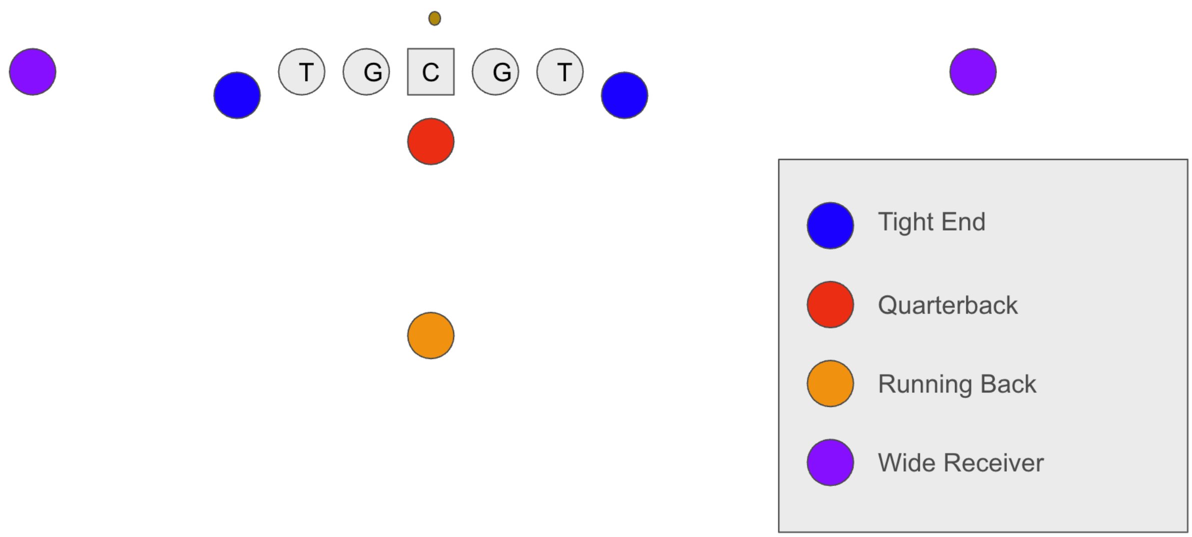

Figure 1

The Data

This competition provided 13 different files of data. 9 of them were tracking data for every play for the first 9 weeks of the 2022 NFL season. The other four were additional data for the plays and players. I only used 2 files from that dataset :

Name	Shape	Description
tracking_week1.csv	(7104700, 18)	Positions of every player and the football for every tenth of a second during a play.
plays.csv	(16124, 50)	Additional metadata for each play found in the tracking files.

Table 1

I won't describe these datasets in full in this paper, but you can see a full list of columns and stats for each column in the 'eda' files. For the purposes of this paper I'll focus on some important findings though. Every frame of every play in the tracking file has exactly 23 entries. This is expected as there are 11 players per side and then a football. Sometimes a team can make a mistake and only field 10 players on a play or field 12 players which results in a penalty. No plays like that existed in this dataset. Some additional histograms can be found in figures 1 and 2 as well.

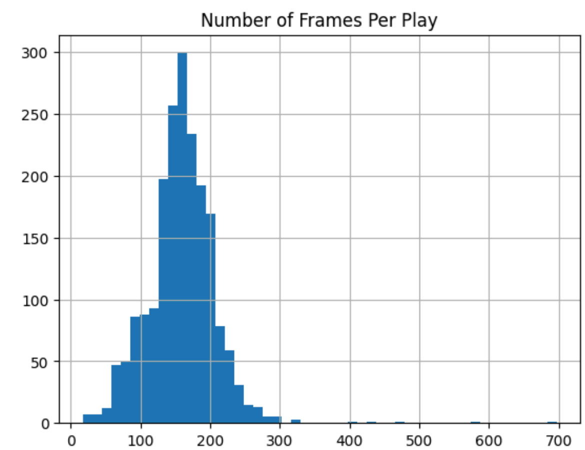

Figure 2

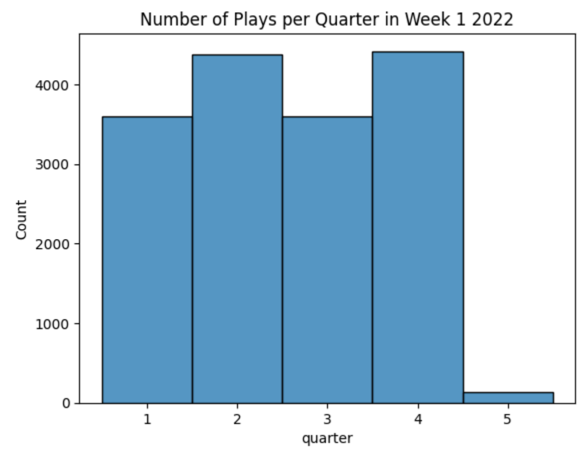

Figure 3

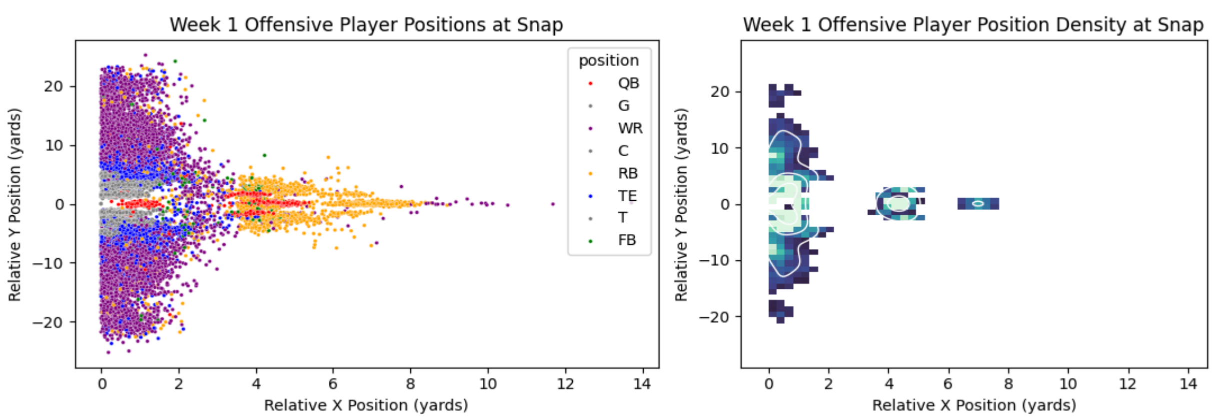

Figure 4: (Left) A plot of the position of all offensive players from week 1 relative to the center. (Right) a density plot showing where any player tends to line up.

Feature Selection (Kitchen Sink)

For every play in the NFL the offense sets the tempo and the defense reacts. Therefore, identifying a formation based on the tracking data seems a little more possible using only offensive players. I also only used the first week of tracking data to speed up iterations and make it easier to understand the visualizations. I tried two different ideas for selecting features for a clustering algorithm. The first (let's call it the "kitchen sink") was to have the relative x and y for every skill position[4] as each feature.

In order to do this I first calculated the relative position of every offensive player to the ball in the frame. The frame I used was the one marked with the event "ball_snap" which means that was the moment that the center snapped the ball to the quarterback. These datasets show the players in the huddle and can last for a very long time so the most useful frame seemed to be the one when I knew all players were in their correct position. See figures 5 and 6 for summaries.

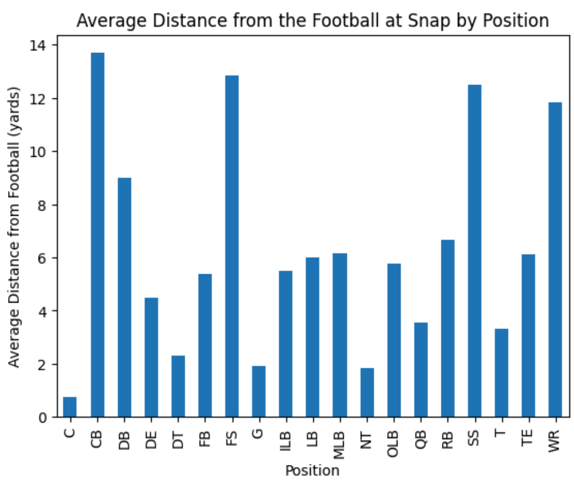

Figure 5

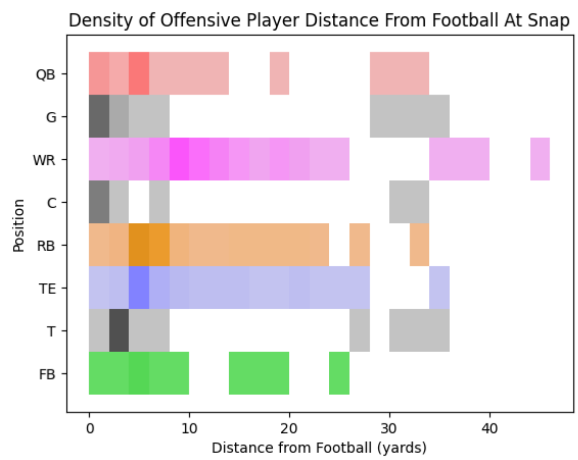

Figure 6

It isn't evident from Figure 3 that anything is wrong. The center is closest to the football (though the value should be 0). The guards (G) and quarterback (QB) are relatively close to the football which is expected. As is the wide receiver (WR) being farther away. However we can see in Figure 4 something unexpected. The center is sometimes 7 or 8 yards away from the ball when it's snapped?! All of the positions have some anomalies when you look closer. The field is 53 yards across its most narrow dimension. The football is placed in between two different hash

[4] Wikipedia contributors. (2025a, February 13). *Skill position*. Wikipedia. https://en.wikipedia.org/wiki/Skill_position

marks that are 70 feet and 9 inches away from their respective sideline. So at most a player could be 30 yards away from the ball when starting the play. Upon further review it was apparent that some of the plays had the ball in an impossible spot at the start of the play (Figure 6). So, instead of measuring from the ball I decided to measure from the center (Figure 7, 8)

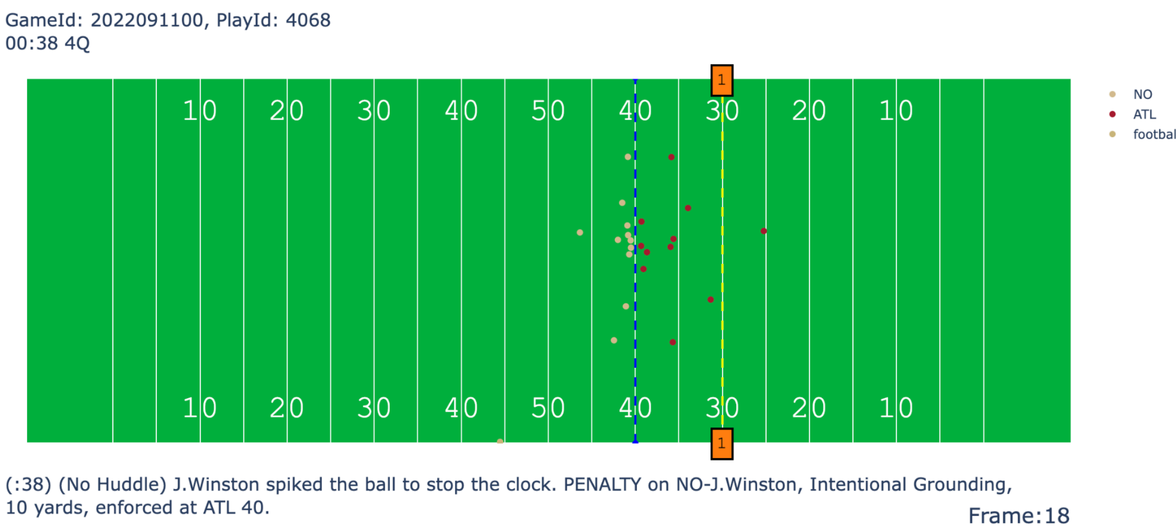

Figure 6

The football can be seen at about the 44 yard line at the bottom of the field at the same time that Jameis Winston is supposedly spiking the ball.

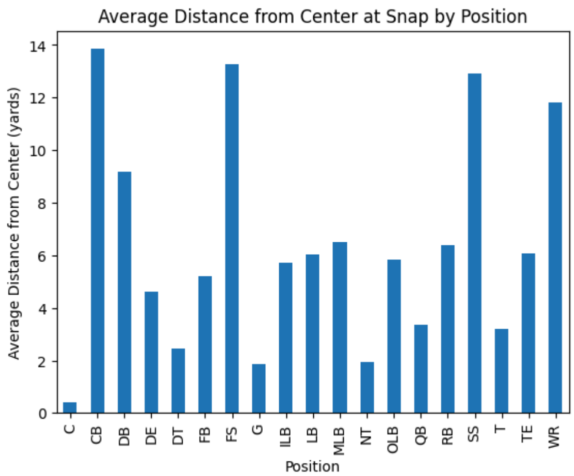

Figure 7

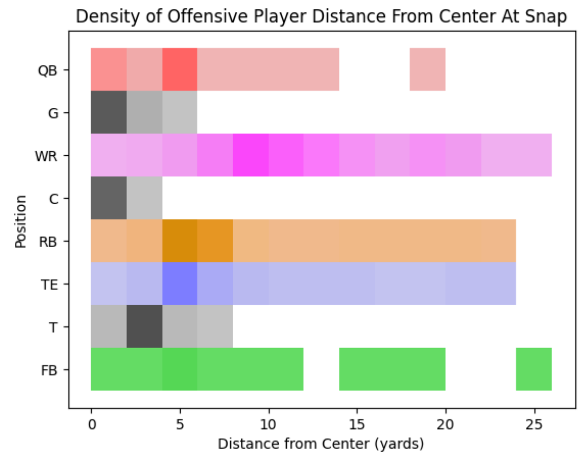

Figure 8

It's still a little odd for the center to have two boxes in the histogram plot (Figure 6). It turns out that in some plays the main position of the player is Center but they were lined up as a guard instead. I do not have supplemental data to show what player lined up at what position on the play so there were a few anomalies like this. A full accounting can be found in Table 2.

Plays where position count is not possible

Position	Anomalous Plays	% of Total (16124 plays)
Center	380	2.3
Quarterback	15	.09
Guards	145	.9

Table 2

I would like to have resolved this issue, and I considered a way of identifying the position the player lined up in, but I decided to see how well the clustering works given these anomalies. I was largely ok with the results so I never came back to this problem.

For the number of possible total positions in a formation I generated a feature for the x and the y value (separately) for each of them. This resulted in 24 total features: 12 positions (5 wide receivers, 1 quarterback, 3 tight ends, 2 running backs, 1 fullback) by 2 values (`relative x`, `relative y`). This would result in a 24 dimensional model where clusters would form where these values were similar. Not every play will have values for all positions. Clustering in general doesn't handle empty values well.

After some experimentation I made two further changes to the final features. The initial tests were resulting in a very poor silhouette score[5] of .25. One concern was that I had 24 features, and also the precision of the relative values were fairly unnecessary. For example a quarterback that is 5 yards behind the center is very similar to a quarterback that is 4 or 7 yards behind the center. So I scaled `x` values to every 5 yards and I scaled `y` values to every 3 yards. In addition I removed the x value for receivers as they generally line up at the line of scrimmage so after scaling they almost all had values of 0 unless they were in motion.[6]

Scaling and reducing the number of parameters did not improve the silhouette score. The best I achieved was a score of .31 with only 3 clusters. I investigated a different method.

Feature Selection (Reduced Method)

I thought about ultimately what the features of a formation are. The positions of the players are certainly important, but in order to reduce the number of features I decided to go with the count of players at each skill position. In addition I used the quarterback's relative `x` position

---

[5] *Selecting the number of clusters with silhouette analysis on KMeans clustering*. (n.d.). Scikit-learn. https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

[6] Swingle, R. (2025, February 24). Football Motion: Pre-Snap movement in the offense - VIQtory Sports. *vIQtory Sports*. https://www.viqtorysports.com/pre-snap-movement-in-the-offense/

as a feature because a quarterback may start right by the center or further back (shotgun). The number of blockers in on a play matters as well. Sometimes a team will use 3 or more “tackles” to provide better blocking for a run play. In the end I only had 4 features: wide receiver count, relative position of the quarterback, tight-end count, and blocker count. This feature set was much easier to reason about and immediately had a much better silhouette score (up to .99 with 50 clusters).

Model Selection

There are a lot of clustering algorithms to choose from, but I experimented with DBSCAN, AgglomerativeClustering, KMeans, and AffinityPropagation from the sklearn package. DBSCAN and AffinityPropagation seemed interesting because they would handle uneven clusters and would also automatically determine the number of clusters instead of expecting a target cluster size. AffinityPropagation failed to yield anything useful to me however. The silhouette score was between -.02 and .02 if it found any clusters at all. It’s clear this shape of data is not good for this algorithm. DBSCAN did a better job of finding clusters with a silhouette score of .93 it found 21 clusters but it had 76 outliers, and small changes in the hyperparameters would lead to no change until a sudden drop off. This may have been a totally valid model, but dealing with 76 outliers was a bit overwhelming.

KMeans and AgglomerativeClustering clustered the plays nearly identically (see ‘ClusterDiff.png’ in the repository). I first iterated through the KMeans algorithm trying a different number of clusters until I discovered a convergence. At 50 clusters KMeans would only give warnings after that so I chose 50 clusters. KMeans and AgglomerativeClustering both gave silhouette scores of .99 suggesting the points are well suited to their cluster. Further hyperparameter tuning did not improve this any.

Results

I then plotted the actual plays together on the same plot per cluster to visualize what that cluster was. The algorithm did a good job. It grouped 3 wide receiver formations with the quarterback in shotgun much like a human would. The outliers (clusters of size 1) were interesting to see as well. In the end this algorithm found 14 unique formations from week 1 of the 2022 season. The full list can be found in "KMeans-one_play_clusters.csv" in the output folder of the repository. But I used that data to track down a few plays from highlight videos on YouTube.

KMeans "Cluster" 23 was a play by the Buffalo Bills where they lined up 2 wide receivers, and extra tackle and a fullback. It looked like it would be a run to the left, but instead Josh Allen looked to throw the ball and then kept it and ran for a first down.[7]

KMeans "Cluster" 34 was a play with no wide receivers, and an extra blocker lined up in the backfield. It looks like a for sure run to the right side but the quarterback keeps it for a touchdown.[8]

---

[7] https://youtu.be/eWPijiN3vGU?feature=shared&t=602
[8] https://youtu.be/w8d70dG4HbA?feature=shared&t=213

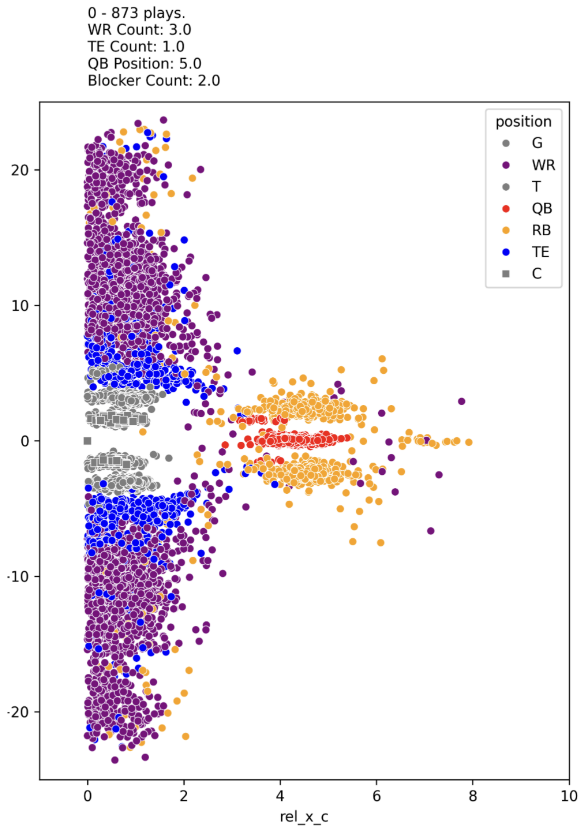

Figure 9

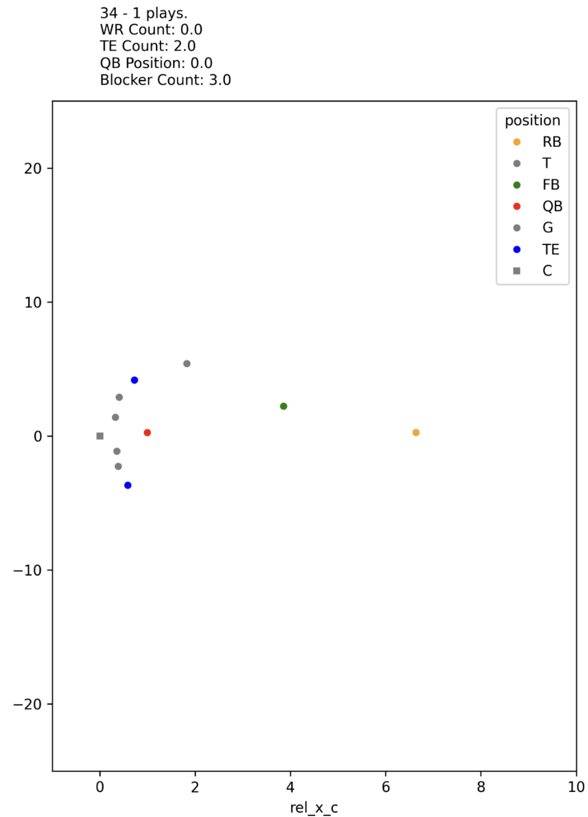

Figure 10

Clusters 0 and 34 showing the variety in clusters found.

I did this exercise for two more videos and each time it resulted in a play that I could actually find in a highlight video on youtube which was surprising. Perhaps these outliers are something that humans would find interesting to review. Coaches may want to see these rare formations that opponents use, or on the other side maybe they want to only review formations that opponents use 80% of the time. Either way, clustering algorithms may be a useful way to accomplish that.