A Strong Passing Game is Key To Winning American Football Games*

Danur Mahendra

27 April 2022

Abstract

A team's success is often attributed to the strength of their roster and ability for star players to win games. Often, viewers fail to consider the statistics and data analysis needed to justify game strategies, which in turn effect how players are used. In particular, certain aspects of the game may have been overvalued by fans and teams when more influential factors could be used to improve a team's chances of winning individual games. In this paper, play-by-play data from the 2021 NFL season was used to analyze factors behind the creation of scoring chances. This paper found that ultimately, passing plays are more likely to generate scoring chances in counterpart to rushing plays. Thus, we conclude that teams ought to prioritize the development and improvement of their ability to pass the football as opposed to their ability to run as a way to extend possessions.

Contents

*Code and data are available at: https://github.com/danurmahendra/Factors-Contributing-to-Scoring-Opportunities-in-the-NFL.

Introduction

Motivation

Sports has always been a subject of interest to me due to its competitive nature and physical health benefits. Professional athletes achieve peak athleticism suited for their respective sport through years of training; and seek to use their physical prowess to create winning plays for their team. Often, viewers attribute team success on the strength of their roster and fail to consider the statistics and data used to aid coaches in decision-making. For example, sabermetrics is the application of advanced statistical analysis used to quantify and predict player performances. It does not evaluate talent based on physical observations but rather, empirical results. Thus, it leads to objectivity and reduces the likelihood of biased decisions. The results could then be used to further improve the team by focusing on the macro aspect of decision making suck as roster construction, player development, and free agency. Similarly, statistics and data analysis can be used to aid coaches and players in pre-game preparation to maximize their likelihood of winning. This paper will look use R (R Core Team (2020)) to analyze the contributing factors in deciding games. Thus, we will look to answer the following questions:

- How are scoring opportunities created?
- What factors contribute to the creation of scoring opportunities?
- How should teams adjust their strategy to maximize scoring opportunities?

Introduction to American Football

The National Football League (NFL) is a professional American football league and considered to be the highest level of play available for its sport. The game is played by two teams consisting of 11 players on each side. Teams alternate possession of the ball and must try to advance the ball towards their opponent's end zone to score a field goal (worth 3 points) or a touchdown (6 points). A touchdown is followed by either a 1-point field goal attempt or a 2-point attempt. The team in possession of the ball, is known as the offense while the opposing side, the defense, must prevent the offence from advancing the ball. The offense is given four down (tries) to try and advance the ball at least ten yards towards their opponent's end zone. Successfully gaining ten yards gives the offense a new set of downs and moves the chains closer towards the end zone. The offensive team ends their possession following an offensive score or a turnover. However, failing to advance ten yards results in a turnover where the defense will now have possession at the current field position. Thus, teams often opt to punt the ball once they reach their fourth down. By doing so, they give up possession but ensures that the opposing team will have a worse starting field position.

Often, games will have key moments capable of shifting the momentum of the game whether it be due to a penalty, completing a crucial play, or injury to a player. Swing moments, are ultimately capable of dictating outcomes of a game. In the NFL, the term "big play" is used to denote high impact plays that exceed its expected value. It does not have a defined threshold on what qualifies as a big play however, a generally accepted bar is a running play that yields more than 12 yards or a passing play that nets the team more than 16 yards. A research done by Seattle Seahawks head coach Pete Caroll found that drives where the offensive team lands a big yardage play results in a score more than 75% of the time (Hsu (2012)). Ultimately, they help extend drives and moves the line of scrimmage closer towards their opponent's end zone which increases scoring opportunities. Accordingly, their offensive game plan revolves around hitting big yardage plays whereas their defense seeks to prevent opponent's big plays. This paper will attempt to analyze further into big plays, determining its cause and likelihood of occurrence.

Type of Play	Number of Occurrence	Total Yards	Minimum Net Yards	25th Percentile	Mean	75th Percentile	Maximum Net Yards
Passing	15,919	116,927	-10.0	0.0	7.3	11.0	99.0
Rushing	11,078	47,216	-13.0	1.0	4.6	6.0	83.

Passing plays tend to result in a greater gain of yardage however, also also associated with more risks. A study conducted by Joel R. Bock found that is almost three times likelier for turnovers to occur as a result of a passing play as opposed to rushing plays (Joel R. Bock (2016)). Consequently, teams may be hesitant calling passing plays to reduce the risk of a turnover.

Data

Data Collection

The data set used was the 2021 NFL play-by-play data and was found on NFLSavant.com (Willman (2022)). The data was, however, collected by the NFL's Football Operations division. Using ball- and player-tracking technology, the Operations division was able to capture real-time data of every players' movement for every play on the field (League (2022)). Specifically, radio frequency identification (RFID) tags were attached to players' shoulder pads as well as the ball to record data. The original data set observed 42,795 plays along with a description for each play such as the home and away team, quarter, time left on the clock, formation, play type, penalties, and yardage gained.

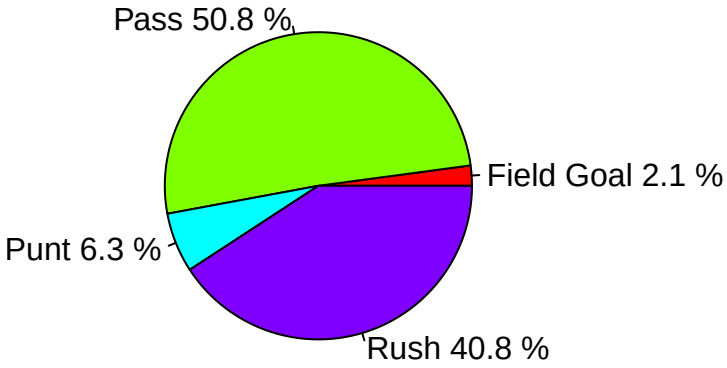
Data Cleaning

For the purpose of this paper, only ordinary plays were considered and those that were situational such as QB kneels, field goal attempts, and extra point conversions were omitted. In football, plays can be designed to accomplish one objective and that only. For example, QB kneels are end-of-game plays usually done by the leading team intended to stall or take playing time off the clock. Doing so meant that the offense would not generate a positive yardage play, but ensures that that the defense would not have a chance at possession thereby securing the win for the offense. I first created a new variable (column) to denote whether the observed play qualified as a big yardage play. A big yardage play is defined as either a passing play that results in a gain of at least 17 yards and running plays that gained at least 13 yards. I then removed any plays that occurred in the overtime as a team's strategy may change depending on whether they received first or second possession. I then omitted plays that resulted in a penalty committed by either team as the yards gained or loss were a result of human error. For instance, an offside is drawn when a defender moves past the line of scrimmage prior to the ball being snapped. Thus, no play had occurred yet and penalties should not qualify as a big yardage play. Lastly, there were four instances where an incorrect data was input was placed under the variable used to categorize pass type. The data did not make sense and appeared to be meant for another variable and thus, was justified in its removal. For example, the following were inputs found under the variable PassType: "INTENDED FOR," and "NOT LISTED." After the cleaning process and sub-setting for plays in ordinary game situations, we finished with 26,997 observations. Cleaning was performed using (Hadley Wickham and Mara Averick and Jennifer Bryan and Winston Chang and Lucy D'Agostino McGowan and Romain François and Garrett Grolemund and Alex Hayes and Lionel Henry and Jim Hester and Max Kuhn and Thomas Lin Pedersen and Evan Miller and Stephan Milton Bache and Kirill Müller and Jeroen Ooms and David Robinson and Dana Paige Seidel and Vitalie Spinu and Kohske Takahashi and Davis Vaughan and Claus Wilke and Kara Woo and Hiroaki Yutani (2019)) and (Wickham et al. (n.d.)). For graphs, charts, and knitting, (Wickham (2016)), (Zhu (2021)) and (Xie (2021)) was used.

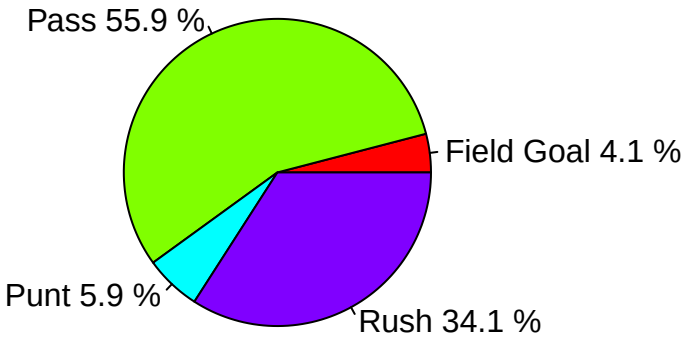
Distribution of Plays By Quarter

Upon further inspection, we observe that passing plays occur at a higher rate than rushing plays. Our observation of 26,997 plays found that roughly 58.9% of all ordinary plays (plays between rushing and passing plays) called results in a passing play. Consequently, it is unlikely to occur by chance and I intend to find out if passing plays benefit teams more than rushing plays.

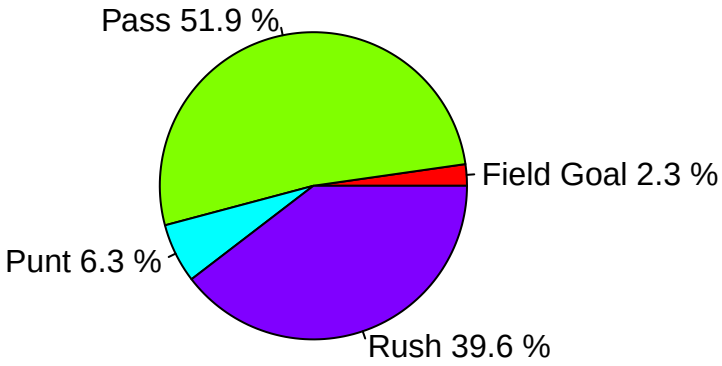
First Quarter Play Distribution



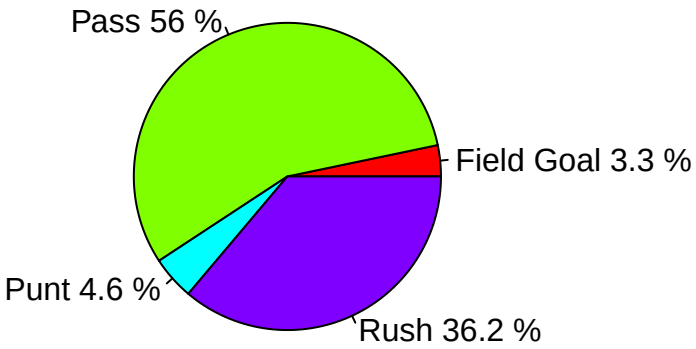
Second Quarter Play Distribution



Third Quarter Play Distribution



Fourth Quarter Play Distribution



Time Period	Passing Plays	Rushing Plays	Big Yardage Plays	>16 Yardage Passing Plays	>12 Yardage Rushing Plays	% of Big Yardage Plays from Pass
1st Quarter	3450	2771	609	429	180	70.4
2nd Quarter	4494	2740	725	575	150	79.3
3rd Quarter	3534	2698	655	474	181	72.4
4th Quarter	4441	2869	716	567	149	79.2
Full Game	15919	11078	2705	2045	660	75.6

Despite passing plays accounting for roughly 59% of all ordinary plays, they account for 75.6% of all big yardage plays. Thus, it makes sense as to explaining why teams favour passing plays over rushing plays. We can then also hypothesize that teams who excel in passing triumphs those whose strength lies in the ability to run the ball. We can put our theory to the test in the following section using models to any relationships between the two play types and generating scoring opportunities.

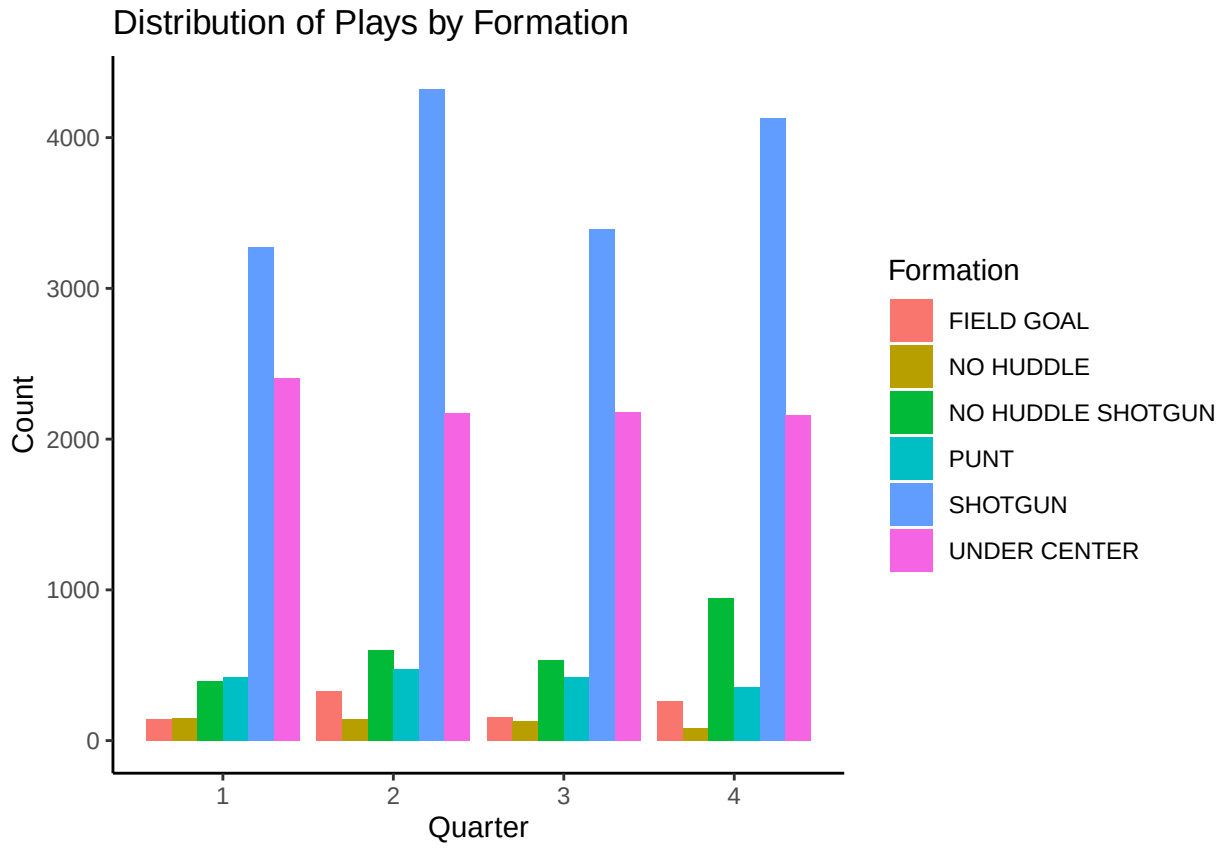


Figure 1: Distribution of Plays by Formation

Methodology

Logistic Linear Regression

This paper's goal is to find the influential factors and conditions which help contribute in big yardage plays. Thus, we would need a binary predictor variable to quantify successful plays and a model that models the probability of an event occurring. Accordingly, a logistic regression model is selected as it can be used to estimate the likelihood of a big yardage play occurring given a set of predictor variables. The predictor variable used for this model is IsBigPlay and denotes whether or not a big yardage play occurred. Ultimately, big yardage plays help extend drives and moves the line of scrimmage closer towards the opponent's end zone thereby increasing scoring chances which contributes to winning. For a more accurate representation, two models were created; one to model the probability of big yardage plays occurring as a result of passing plays, and the other via rushing plays. Using R (R Core Team (2020)), I will use a logistic regression model which takes the form:

Model Specifics

Using R (R Core Team 2020), I will use a logistic regression model which takes the form:

$$log(p/(1-p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

The variable p is used to denote the probability of the event occurring which in this case, is either a gain of at least 17 yards from a passing play or a gain of at least 13 yards as a result of a rushing play. $\beta_2, \beta_3, and \beta_4$ are correlation coefficients that provides appropriate weighted values to each independent variables. X_1 is the direction the play is headed towards. X_2 is the formation of the offense at the time of the snap. X_3 is the distance in yards required to achieve a first down. X_4 is the time left on the clock before the quarter ends. The factors, or independent variables, used to predict event occurrences vary between play types as it is impossible to join certain factors together. For example, there cannot be an input for rush direction if the play called was a passing play. Consequently, a model was created for each type of play that can later be compared to determine which best produces big yardage plays. More specifically, the model for passing plays can be modelled by:

$$log(p/(1-p)) = \beta_0 + \beta_1 PassType + \beta_2 Formation + \beta_3 ToGo + \beta_4 Minute + \epsilon$$

and for running plays,

$$log(p/(1-p)) = \beta_0 + \beta_1 RushDirection + \beta_2 Formation + \beta_3 ToGo + \beta_4 Minute + \epsilon$$

As we can see, the graphs in Figures 2 and 3 appear right-skewed although its degree or severity may vary. Rushing yards have a higher concentration of plays towards the right-side spectrum of the graph. As a result, it means that their plays tend to result in a lesser gain of yards. In contrast, the graph depicting passing yardage have long tails; thereby has a higher frequency to generate big yardage plays. Consequently, passing plays have a higher likelihood of generating big yardage.

Majority of yardage gained were a result of short passes. However, big yardage plays were often a result of deep passes and more specifically deep middle passes. Thus, the ability to generate and convert deep passes are significant factors that contribute to winning games.

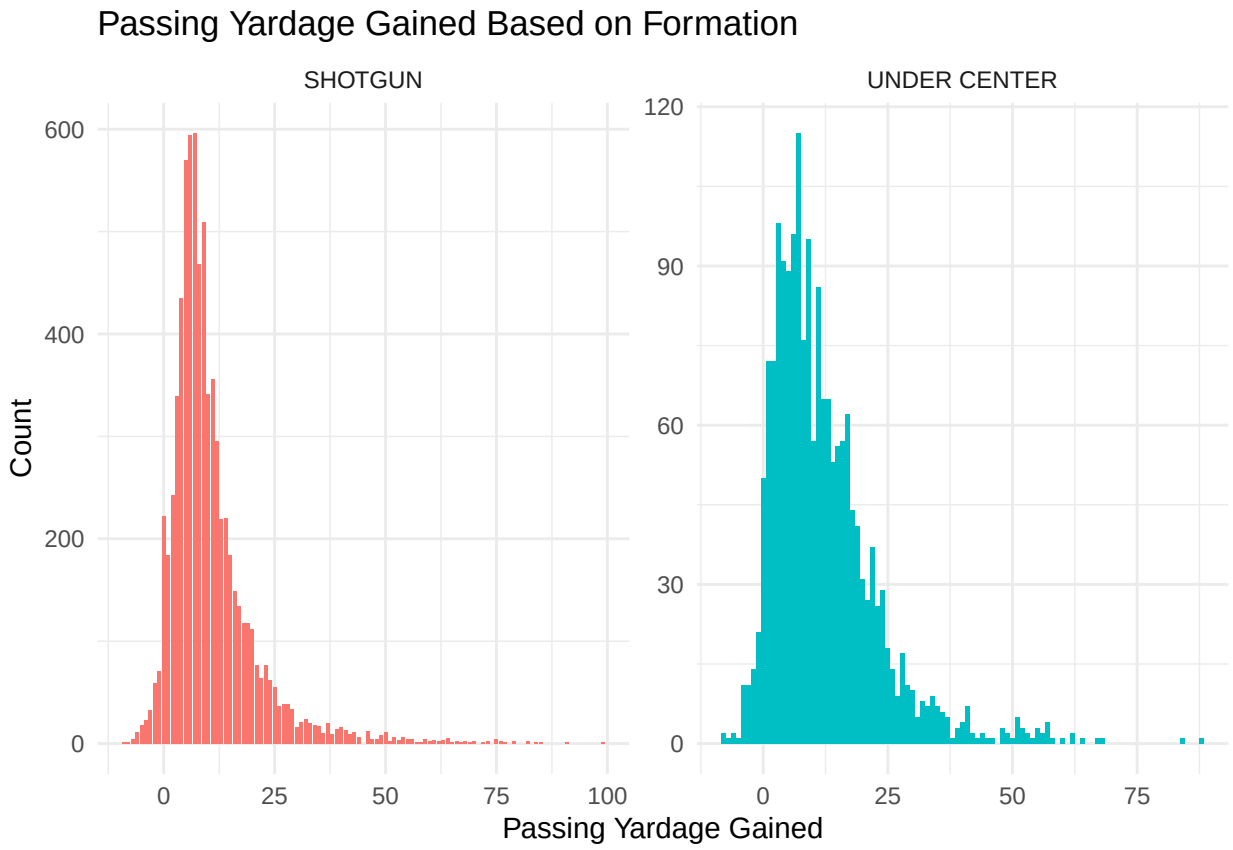


Figure 2: Passing Yardage Gained Based on Snap Formation

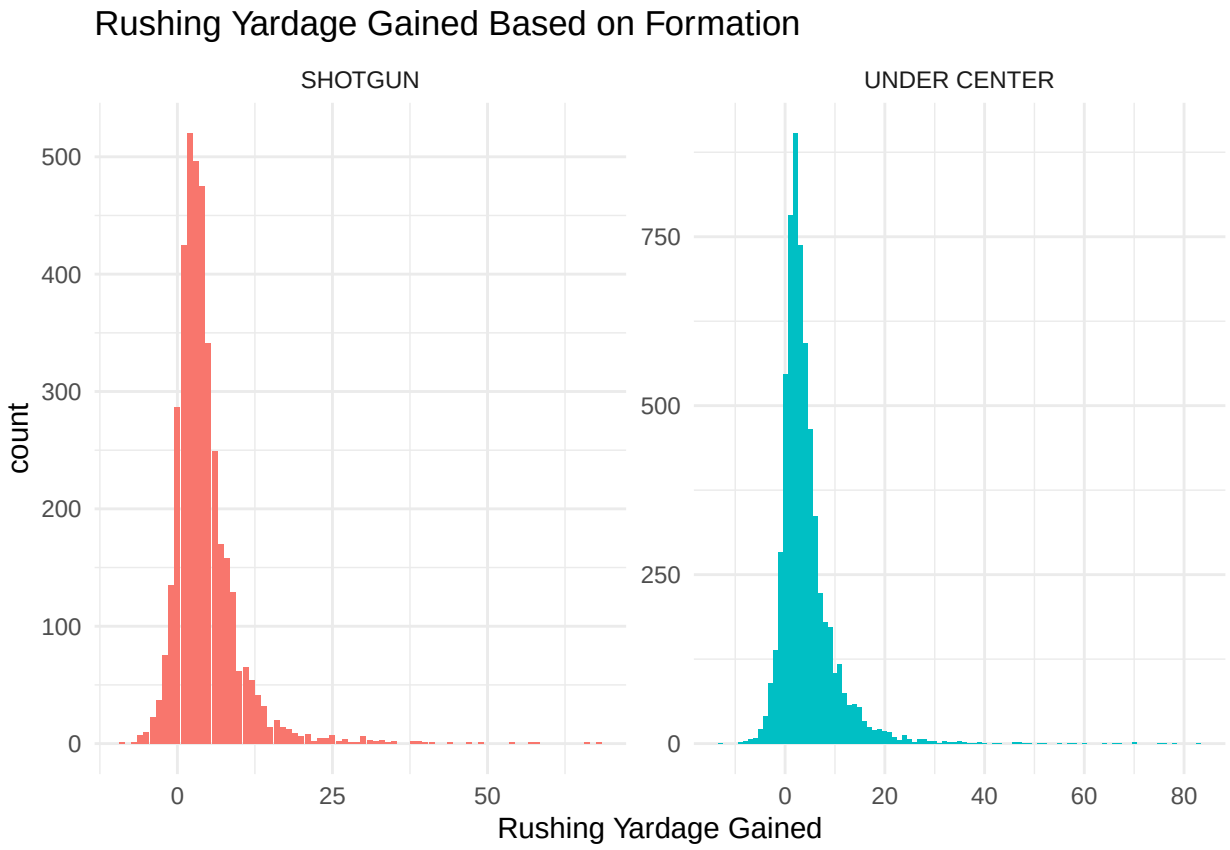


Figure 3: Rushing Yardage Gained Based on Snap Formation

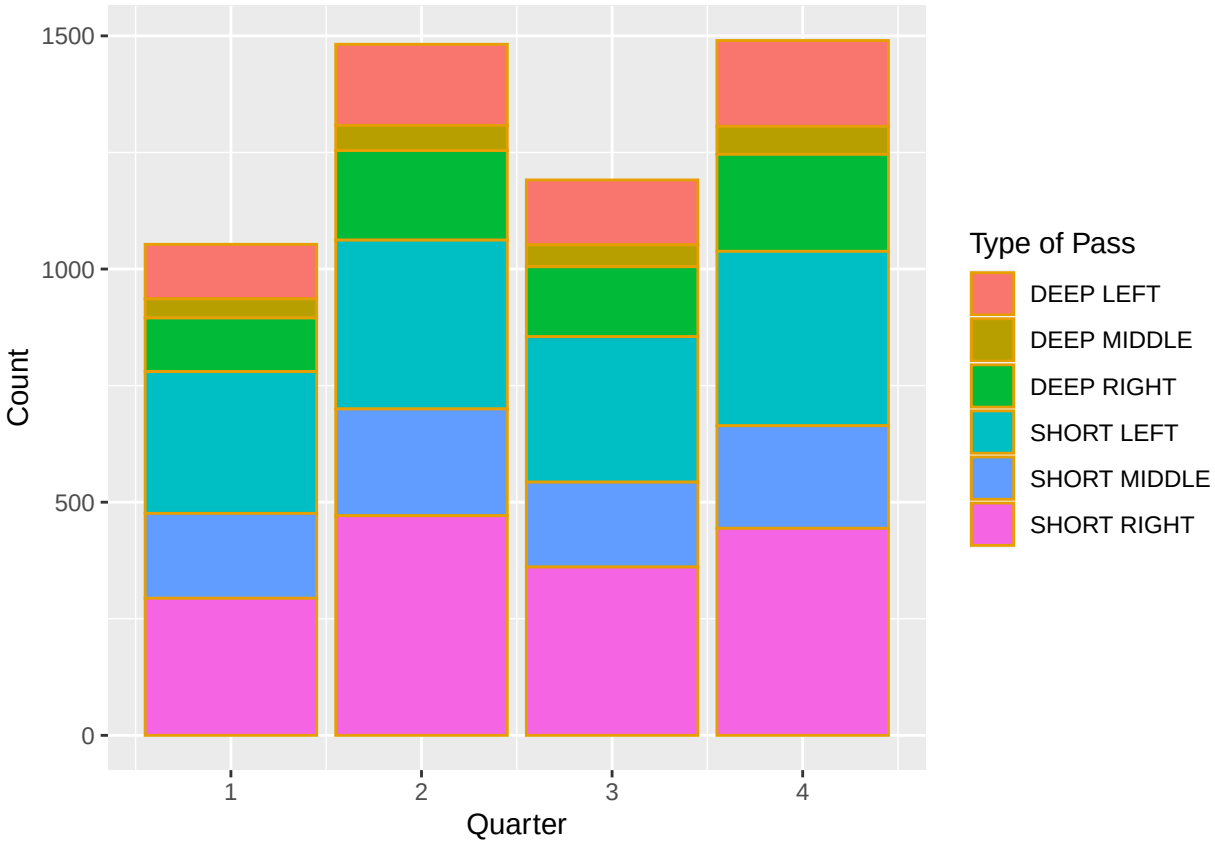


Figure 4: Total Composition of Yardage Based on Type of Passs

Results

Model Output

Using logistic regression, we were able to obtain the following output for our two models. The model with respect to passing plays,

Variable	Category (variable)	Estimate	Std. Error	z-value	P value
Intercept		-18.74	219.254	-0.085	0.93187
PassType	DEEP LEFT	2.38	0.077	30.561	< 2e-16
PassType	DEEP MIDDLE	2.94	0.101	29.054	< 2e-16
PassType	DEEP RIGHT	2.38	0.073	30.945	< 2e-16
PassType	SHORT LEFT	0.29	0.077	3.964	7.36e-05
PassType	SHORT MIDDLE	0.72	0.077	9.278	< 2e-16
PassType	SHORT RIGHT	0.16	0.073	2.269	0.02326
Formation	NO HUDDLE	15.71	219.254	0.072	0.94286
Formation	NO HUDDLE SHOTGUN	15.54	219.254	0.071	0.94347
Formation	PUNT	-0.05	271.010	0.000	0.99986
Formation	SHOTGUN	15.57	219.254	0.071	0.94339
Formation	UNDER CENTER	15.91	219.254	0.073	0.94216
ToGo		0.015	0.005	2.632	0.00848
Minute		0.013	0.005	2.685	0.00726

and with respect to rushing plays,

Variable	Category (variable)	Estimate	Std. Error	z-value	P value
Intercept		-18.75	219.221	-0.086	0.931814
RushDirection	CENTER	-1.45	0.101	-14.279	< 2e-16
RushDirection	LEFT END	-0.47	0.099	-4.786	1.70e-06
RushDirection	LEFT GUARD	-1.21	0.128	-9.434	< 2e-16
RushDirection	LEFT TACKLE	-0.96	0.117	-8.298	< 2e-16
RushDirection	RIGHT END	-0.49	0.104	-4.699	2.62e-06
RushDirection	RIGHT GUARD	-1.36	0.134	-10.118	< 2e-16
RushDirection	RIGHT TACKLE	-0.84	0.112	-7.495	6.63e-14
Formation	NO HUDDLE	16.74	219.221	0.076	0.939130
Formation	NO HUDDLE SHOTGUN	16.56	219.221	0.076	0.939753
Formation	PUNT	-0.04	270.951	0.001	0.999855
Formation	SHOTGUN	16.53	219.221	0.075	0.939862
Formation	UNDER CENTER	16.91	219.221	0.077	0.938488
ToGo		0.02	0.005	3.603	0.000314
Minute		0.01	0.005	1.957	0.050389

Inference

According to the model's output, the offense has the highest likelihood of generating a big play by throwing deep middle passes while under center. As for rushing plays, the model suggests a rush towards the left end from an under center formation. Our model also suggests that extending the clock and distance for a first down as far as possible increases opportunity. Admittedly, it is extremely unlikely for a team to find themselves needing 99 yards for a first down at the 15th minute. However, it does make sense that it maximized the likelihood to generate a big play. I will address this issue in the following section. Accordingly, we see

that formation has a high p-value (<.900) for both passing and rushing plays. Thus, it is inconsequential as a factor as no meaningful relationship can be drawn. However, the low p-values from the type of play, distance to a first down, and time left on the clock suggests that it plays a factor in creating big yardage plays.

Overall, we can generally conclude that teams who excel in passing the football generate more big yardage plays than teams who do not, or rely on other methods to advance the line of scrimmage. Generating big yardage plays generally increase scoring opportunities by advancing the line of scrimmage closer to the opponent's end zone. Thus, being able to consistently do that would really put a team in winning positions.

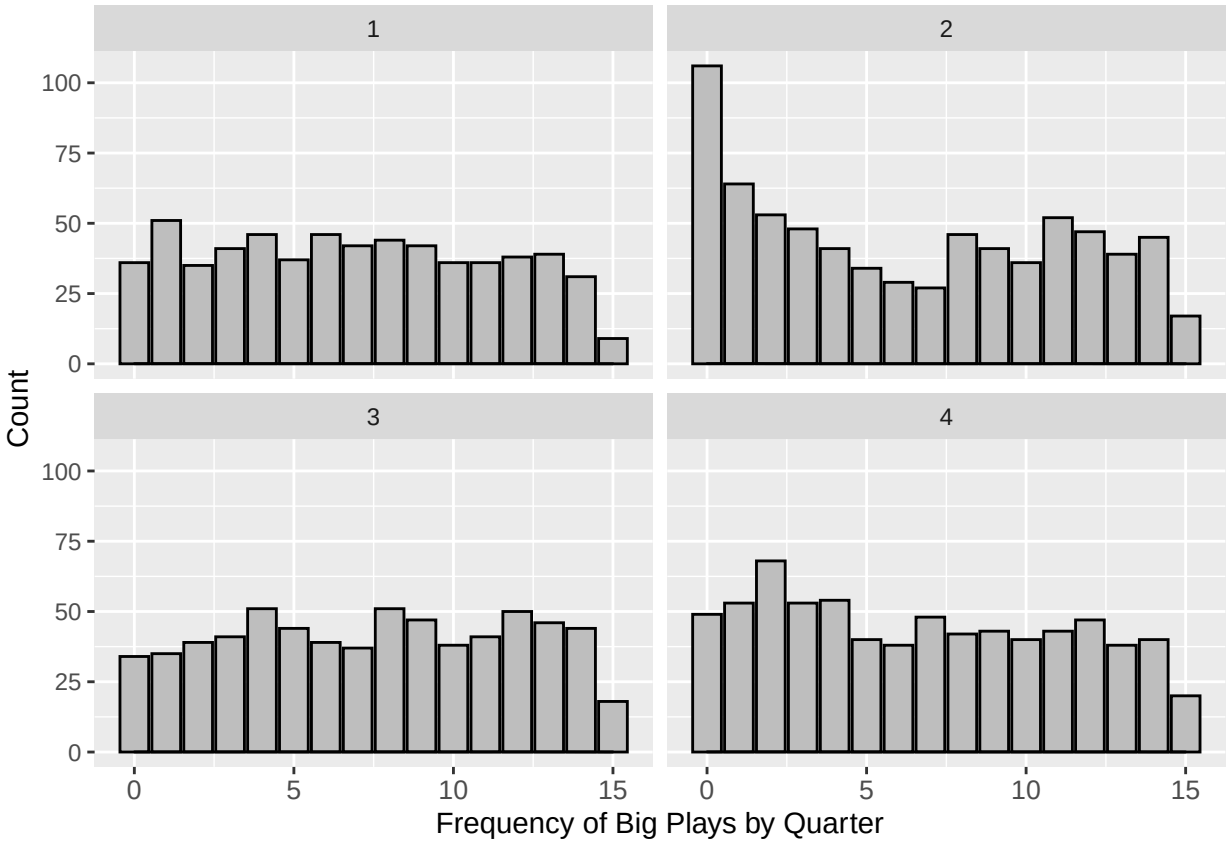
Limitations and Weaknesses

Limited in Reality

Our model suggests that teams have the highest likelihood of generating a big yardage play given that they are as far away from the first down marker as possible, and in the final minute of play. Although it does make sense, it ultimately does not increase a team's likelihood of winning the game. Consider this scenario, a team finds itself in a 4th & 99 spot; meaning they need 99 yards to extend the drive. Anything less than 99 yards is deemed a success for the defense as it results in a turnover. The defense knows this so accordingly, they are likely to shift their resources on preventing plays that could result in a 99-yard gains (ie. hail-mary plays). In particular, they may dedicate extra safeties and cornerbacks in exchange for linemen to limit the possibility of a blown coverage. As a result, the defensive line and area near the line of scrimmage is vacated as extra manpower is dedicated towards the downfield. To better explain the situation, consider this video (https://www.youtube.com/watch?v=Bd3bJ9gvD44) (Highlights (2017)) at 0:01-0:02, we see an open short field as the defense only rushed 3 out of 11 defenders towards the QB. This opens the possibility of a short pass with yards after catch (YAC) to extend the pass. Such pass, while it may look good on the box score, is meaningless as time would have ran out and thus, automatically ends the drive. However, later on at the 0:09 mark, we see the remaining 8 out of 11 defenders downfield defending the end zone as it is the only meaningful completion the offense can throw to affect the game.

A solution to fix this issue could be to constraint the distance to go variable as it fails to take into account specific game situations.

Factors Beyond Team's Control



We see that that the start of the second quarter sees an increased likelihood of a big play occurring. However, one must consider the possibility that the factors causing the increased likelihood may be beyond a team's control. After every quarter, teams take a short break in addition to switching sides of the field. After the first and third quarter, a 2-minute break is given whereas a 12-minute break follows at half-time (after the second quarter). Accordingly, this gives teams a short period of time where play is halted. In the NFL, a time-out gives teams 1 minute to halt play and use that time to communicate with their coaches. They may discuss strategies, make player substitutions, analyze the opponent's weaknesses, or inspire morale. It is also important to note that the defense are also given the same privileges as the offense. Post-quarter breaks, however, gives at least double the time one would receive from a timeout. Therefore, we can infer that the time-outs or simply put, a short break in play, benefits the offensive team more than the defending team. Thus, we can infer that teams who begin the second quarter with possession of the ball may see an additional advantage that would not have been given for any other reason but game time.

All Teams Were Weighted Equally

Unfortunately, the data set obtained did not consider or rank the varying roster strength between teams. For instance, the 1st-seeded team is undoubtedly more talented than the 30th-seeded team. Thus, it is reasonable to assume they may have access to a wider variety of play calls and likelihood of achieving success. I would have loved if for the data set to include the team's rank based on certain team sets such as their offensive ranking, rank in total passing yards, rushing yards, and defensive ranking. I could then use their rank statistics as part of my predictor variable to find a relationship

Appendix

Data Sheet

The template to the data sheet used can be found (Gebru et al. (2021))

Motivation

1. *For what purpose was the data set created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - The data set was created for the purpose of tracking NFL plays in real-time. Using player- and ball-tracking technology, teams could use data to make precise adjustments in their strategies. Unfortunately, we were unable to quantify and weight a team's strength in relation to their total yardage. For example, the 1st-seeded team is expected to generate better, and higher scoring opportunities compared to a team at the bottom of the standings. Thus, some plays may work for them which otherwise would not for other teams.
2. *Who created the data set (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The National Football League's Football Operations division collected and created the data set for the League.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
    - The National Football League
4. *Any other comments?*
    - N/A

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
    - They represent the plays ran by an NFL team. More specifically, it includes game state and the results. Yes, there are multiple instances where a certain play is ran multiple times from the same yard line, distance to first down, or formation. Results, however, varied between those instances.
2. *How many instances are there in total (of each type, if appropriate)?*
    - TBD
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
    - The data set is a play-by-play data all plays in the 2021 NFL season. Thus, it is not a sample but rather, population of NFL plays.
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
    - Each instance consists of game date, team name, play description, and states of the game relevant to play-calling.

5. *Is there a label or target associated with each instance? If so, please provide a description.*
    - N/A
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*
    - Yes, varying plays may not fit in certain variables and thus, were left with no input. It is intended and not as a result of mechanical or human error. For example, a passing play cannot have a rushing direction, or rushing yards gained. However, it is explained by variables related to passing statistics.
7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
    - None available
8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
    - None available
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
    - Yes, there were several errors where the input did not make sense. Instead, it appeared that the input was intended for other variables. For example, under "PassType" I found several out of place inputs such as "INTENDED TO" or "MIDDLE TO" with no other instances of it occuring. Thus, it led me to believe it was an error.
10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
    - No, the dataset was provided by NFLSavant.com however, it was collected by the National Football League and its Football Operations division. NFLSavant.com is not an NFL affiliate.
11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
    - No.
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
    - No.
13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
    - No.
14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
    - Yes, player information is recorded such as jersey number, first name initial, and last name (ie. 8-K.COUSINS)

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - No.

16. *Any other comments?*

    - N/A.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - Data was collected in real-time during each play. It was then published afterwards on various live-feed sites via API. The data is validated as the observation can be witnessed by numerous media and sports outlets.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - Radio frequency identification tags were placed in players' shoulder pads, the football, and various positions of the field to track a subject's movemand and location.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - The data set consists of the full population

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - The NFL's Football Operations staff was involved for data collection. I would assume they were paid atleast minimum wage based on their state of employment.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

    - The data was collected for the 2021 NFL season thus, between September 10th and January 3rd of 2021.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - None available

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - I obtained it from a third-party website, NFLSavant.com

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

- The data was publicly available thus, no notification was necessary.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
    - Yes.
10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
    - Yes, the NFL has a terms and conditions that all parties must follow.
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - Teams have conducted their own analysis using this data set or a similar data set however, they may want to keep it confidential as to prevent the public and opponents to determine their game strategy.
12. *Any other comments?*
    - N/A

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
    - Yes, several missing or incorrect values had to be fixed.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
    - Yes.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - Yes, data was cleaned manually using R and its functions.
4. *Any other comments?*
    - N/A

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - Personally, no. However, other people may have used it.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
    - None that I was able to find
3. *What (other) tasks could the dataset be used for?*

- It could be used to find trends in NFL plays such as highest scoring quarter, likelihood of converting points off turnovers, etc.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

   - No.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

   - No.

6. *Any other comments?*

   - N/A.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

   - Yes, the data is publicly available and can be accessed by API.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

   - Data was distributed through API

3. *When will the dataset be distributed?*

   - The plays were available immediately however, this particular data set was finalized at the conclusion of the 2021 NFL season.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

   - No.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

   - No.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

   - No.

7. *Any other comments?*

   - N/A

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

- The data set has been finalized therefore, no changes are necessary to support the data set. However, multiple websites may host the data set. This data set was in particular, hosted by NFLSavant.com

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - The owner, Daren Willman, can be contacted by Twitter handle darenw or email darenw@gmail.com
3. *Is there an erratum? If so, please provide a link or other access point.*
   - No.
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
   - No.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
   - No, all NFL players agree to take part in the League's data collection system.
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
   - Yes, data from previous years' play-by-play data sets are available.
7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
   - No, there is nothing to be added on as the plays/observations have been finalized.
8. *Any other comments?*
   - N/A

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Hadley Wickham and Mara Averick and Jennifer Bryan and Winston Chang and Lucy D'Agostino McGowan and Romain François and Garrett Grolemund and Alex Hayes and Lionel Henry and Jim Hester and Max Kuhn and Thomas Lin Pedersen and Evan Miller and Stephan Milton Bache and Kirill Müller and Jeroen Ooms and David Robinson and Dana Paige Seidel and Vitalie Spinu and Kohske Takahashi and Davis Vaughan and Claus Wilke and Kara Woo and Hiroaki Yutani. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43). https://doi.org/10.21105/joss.01686.

Highlights, Rookie. 2017. "Baker Mayfield INSANE Hail Mary Touchdown Pass to Donovan Peoples-Jones | Cardinals Vs Browns." youtube. 2017. https://www.youtube.com/watch?v=Bd3bJ9gvD44.

Hsu, Davis. 2012. "Pete Carroll's Defensive Priorities, Part i." 2012. https://www.fieldgulls.com/2012/11/8/3618798/seahawks-pete-carroll-defensive-priorities-part-i.

Joel R. Bock. 2016. "Empirical Prediction of Turnovers in NFL Football." *Sports (Basel)* 5 (1). https://doi.org/10.3390/sports5010001.

League, National Football. 2022. "NFL Next Gen Stats." 2022. https://operations.nfl.com/gameday/technology/nfl-next-gen-stats/#:~:text=Balancing%20innovation%20with%20tradition%2C%20the,play%2C%20anywhere%20on%20the%20field.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. n.d. *Dplyr: A Grammar of Data Manipulation*.

Willman, Daren. 2022. "Can i Have Your Data?" 2022. http://nflsavant.com/about.php.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://cran.r-project.org/web/packages/knitr/index.html.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://cran.r-project.org/web/packages/kableExtra/index.html.