Final Project Phase 2: NFL Big Data Bowl 2026 Preliminary Exploration

Josh Gaughan

October 25th, 2025

Research Question 1

Motivation

1. **How many yards of average separation from a defender does a receiver need to have in order to gain a certain amount of yards?**
2. The data sources that I will be using for this question include all three of the data sets provided by the NFL. These include the input (pre-pass) data set, the output (post-pass) data set and the supplementary data set. These data sets will support the question by providing the field position coordinates for each player, which allows for the calculation of separation using the Euclidean distance formula. The supplementary data set will also provide the yards gained metric as well.
3. This research question can inform coaching or other staff members how much separation defenders need to achieve in order to gain a certain amount of yards. As part of successfully answering this question in a full data bowl submission, I intend to also provide a way for coaches to see which receiver routes are more likely to produce successful gains so they can choose these routes when trying to achieve certain yards.

Methods

1. In order to answer this question, it is important to have field position coordinates for receivers and defenders. We will utilize the input and output data provided by the NFL for game, play and player identification. This data will also be the source of our field position. We will also use the supplementary dataset provided by the NFL in order to get yards gained. To check data integrity, I first tried to find if any duplicate plays existed in any of the data sets. For the input and output data, I did this by trying to find if the count of each unique combination of game_id, play_id, frame_id and nfl_id (unique player identification code) was more than one. For the supplementary data, I checked to see if the count of each unique combination of game_id and play_id was more than one. For each data set, there were no duplicates. I also checked for missingness in all of the columns in each data set. There were no NAs in the input or output data and NAs in columns we are not going to use in the supplementary data set. Finally, I completed range checks to see if the x and y coordinates for each play were within range. There were three

rows (out of 562,936 in the output data) that were slightly outside of range which is not a concern for our analysis.

2. Please see sections 3 and 4 of the accompanying R markdown or knitted HTML file for EDA and models. I looked at average pre and post throw frame times per game. I also looked at the frequency of offensive route and defensive coverage types to gain an understanding of team tendencies.

Results

1. Please see section 4.10 in the accompanying Rmd file for the corresponding plot for question 1. The plot shows the generalized additive model's predictions for how much average yardage of separation a receiver needs while the ball is in the air in order for the team to gain a specific amount of yardage in a play.
2. According to the plot there is a sharper increase in separation required to achieve small amounts of yardage before flattening out. According to the model, to achieve 10 yards of gain, the receiver must average about 4 yards of average separation while the ball is in the air during a 10 yards pass. The model also indicates that, to gain 20 yards, the receiver must average about 8 yards of separation while the ball is in the air.
3. This model and the predictions it produces give an idea of the yardage of separation required for the team to gain a certain amount of yardage. The model can give coaches an idea of receiver routes they want to run based on the amount of yardage they need to gain.

Discussion

1. The key limitation of this project proposal is my individual knowledge of football. I believe this may provide helpful insights but I think more experienced and knowledgeable individuals will be able to weigh in on how helpful or unhelpful this is.
2. For a full big data bowl submission, I believe that there would need to be a review of the model used and the methods I applied. A generalized additive model may or may not have the best performance for predictions in this specific scenario. I don't think I would ditch the idea altogether but I don't know if it's really worth consideration for a full submission.
3. If anyone else is doing analysis on defender separation (which I assume people are), then collaborations may produce more comprehensive and helpful results.

Research Question 2

Motivation

1. **What is the relationship between team-wide offensive separation from defenders and win probability?**
2. For this question I will use all three of the data sets provided by the NFL. As in question 1, I will utilize the input (pre-pass), output (post-pass) and supplementary datasets. The input and output data sets will provide player coordinates for the Euclidean distance calculation for defender separation. The supplementary dataset will provide the win probability metric.
3. This question can provide a greater understanding of how certain amounts of team-wide offensive separation influence a team's winning probability. The intention is for coaching staff to have the ability to decide what play they want to run and how certain plays and the separation they create can increase their win probability the most.

Methods

1. To answer this question, we will want to utilize the provided input and output data in order to get game, play, player identification in addition to field position coordinates for all offensive and defensive players. We will also use the provided supplementary dataset for the win probability data. The same data integrity checks were done for question 2 as in question 1 since they use the same data sets. This includes the duplicate plays check to find any repeat rows in the input, output and supplementary data sets. I also completed the missingness check for all columns across the data sets. Finally, I checked to see if the winning probability metric was within the expected range and it was.
2. Please see sections 5 and 6 of the accompanying R markdown or knitted HTML file for EDA and models. I checked the minimum, maximum and mean frame times in the input dataset. I also looked at the distribution of player roles as well as the distribution of player positions. Finally, I looked at the average speed and acceleration of each player position.

Results

1. Please see section 6 in the accompanying Rmd file for relevant output. The model summaries for a linear and generalized additive model and plot seem to indicate that there isn't a significant relationship between average team separation and win probability added.

2. The model and plot showed that correlation between average team separation and win probability added was low and pretty flat. When adding covariates to the model, the model performance improves which may indicate that average team separation isn't a super strong predictor of win probability added.
3. The results of this analysis are unhelpful when it comes to the question presented. This could be for a multitude of reasons but if the findings somewhat resemble correctness then the question may not have a very insightful answer.

Discussion

1. Confounding factors may be a limitation for this question. There may be additional factors that are affecting the outcome that need to be explored before moving any further with this question.
2. I do not believe this submission is as strong or helpful as question 1. I had some difficulty with organizing the data and so there may be some further organization that could be done to get a cleaner answer. I don't think I would continue with this submission.
3. Consulting other students who understand football more and who understand modeling a bit more may be a helpful collaboration for answering this question.

Works Cited

NFL 2026 Big Data Bowl Data Sets

Michael Lopez, Tom Bliss, Ally Blake, and Addison Howard. NFL Big Data Bowl 2026 - Analytics. https://kaggle.com/competitions/nfl-big-data-bowl-2026-analytics, 2025. Kaggle.