Unraveling the Dynamics of the NFL Passing Game: The Impact of Downfield Yardage on Passing Play Success

Stephen K. Skasko

University of Pittsburgh

STAT 1341: SPORTS ANALYTICS

Dr. Nelson

October 29, 2023

Abstract

This research delves into the dynamics of the National Football League (NFL) passing game, focusing on key variables such as yards_gained, air_yards, and passing_yards. By employing a multiple linear regression model and comprehensive statistical analyses, it uncovered crucial insights into the relationships between these key variables, shedding light on the strategic and performance aspects of NFL teams during gameplays. The findings emphasize the significance of downfield passing and overall yardage gains in shaping successful passing plays, showing valuable implications for coaches, players, and management. The study identifies potential areas for future research, highlighting the need to incorporate defensive strategies, player-specific dynamics, external factors, and advanced statistical techniques to capture the nature of team performances in the NFL.

Unraveling the Dynamics of the NFL Passing Game: The Impact of Downfield Yardage on Passing Play Succuss

Introduction

The passing game in the National Football League (NFL) is a crucial component of offensive strategy. Teams and sports statisticians need to understand the intricate dynamics of this game to optimize their performance and gain a competitive edge. The objective of this paper is to analyze key variables, such as yards_gained, air_yards, and passing_yards, uncovering trends and patterns in the NFL passing game. By examining the relationships between these variables, the aim is to provide valuable insights that can assist in strategic decision-making within the league, helping to unravel the dynamics of the NFL passing game for the impact of downfield yardage on passing play success.

Data Collection

The dataset used for the analysis was scrapping nflfastR from 2013 to 2022. Before the analysis, data cleaning is used to ensure data accuracy and consistency. This includes the removal of incomplete observations and extraneous variables, which consisted of passing_yards, air_yards, and yards_gained.

Descriptive Statistics

The most valuable variables identified are yards_gained, air_yards, and passing_yards. These variables were crucial for understanding the downfield passing in NFL games. The variable yards_gained represents the total yards gained in a play, while air_yards shows the distance the football is thrown down the field. Passing_yards is the number of yards gained from passing the ball and is the response variable. The variables' mean, standard deviation, and median are significant. Yards gained have an average mean of 4.03, a standard deviation of 7.58, and a median of 1.00. Air yards have an average of 7.74, a standard deviation of 6.29, and a median of 7.74 while passing yards have an average mean of 10.93, a standard deviation of 4.85, and a median of 10.93. With this, a histogram plot compares yards_gained with frequency and a boxplot for passing_yards with air_yards, which helps to understand the relationship between key variables. The histogram plot shows that most yards the NFL teams go for are under 20 yards while throwing over 20 yards is not frequent. The boxplot shows small interquartile (IQR) data points, which indicates low variability for data between the ranges of -30 to 65 air yards gained. These two plots show the range of air yards gained from 2013 to 2022 to indicate potential trends or patterns in the data.

Mean	Average value of a dataset
Standard Deviation	Measure of data spread around the mean
Median	Middle value in a dataset
Correlation	Statistical relationship between two variables
Correlation Coefficient	Numerical measure of the strength and direction of a relationship
Interquartile Range (IQR)	Range between the 25th and 75th percentiles of a dataset

Table 1. Key terms used in the descriptive statistics section

Inferential Statistics

Additionally, the model used is a multiple linear regression to help understand the relationship between the response variables, passing_yards, and the predictor variables, air_yards and yards_gained. The linear regression model showed significance with a low p-value (<2e-16) for all three predictors, while the coefficient estimates for air_yards and yards_gained are 0.295598 and 0.357092. These results suggest an increase in air_yards being associated with around 0.30 increases in passing yards, and in yards_gain with an approximate 0.36 increase in passing yards, all else being equal. The R-squared value is 0.5493, showing a 55% variability in the response variable of passing_yards, indicating a moderate-to-strong fit. The adjusted R-squared (0.5493) is close to the R-squared (0.5493) value, suggesting that predictors did not significantly affect the model's overall fit. The residual standard error is 3.256, and the residual mean (-6.70e-17) is relatively close to zero to further suggest that the model is capturing most of the variability in the data. The residual distribution shows a minimum value of -29.270 and a maximum of 44.567, indicating that there may be some outliers or influential data points in the data set. The analysis of the variance table shows the significant impact of both air_yards and yards_gained on the variability in passing_yards, with a low p-value (<2.2e-16) for both predictors, reinforcing the importance of these two variables in predicting the response variable. The cross-validated result supports the model's performance, with the root mean squared error (RMSE) of 3.256485, the R-squared of 0.5486726, and the mean absolute error (MAE) of 2.248118. This shows the model's ability to generalize data with reliability and predictive power.

These plots provide insights into the performance of our regression model. However, the residual plots indicate that the model may not fully meet the assumptions of linearity and consistent variance, suggesting the need to refine the model to capture better data patterns. The Q-Q residual plot's deviation from normality shows that some data points may have a greater impact on our results than others. Additionally, the presence of influential data points in the leverage plot highlights the importance of closely examining these points to understand their effects. The VIF plot's demonstration of multicollinearity warns us that some of the predictor variables may be highly correlated, which could affect the model's results.

P-value	A measure of the strength of evidence against the null hypothesis
Multiple Linear Regression Model	Statistical model that examines the relationship between multiple independent variables and a dependent variable.
R-squared	Proportion of the variance in the dependent variable that is predictable from the independent variables.
Residual Standard Error	Measure of the variation in the residuals in the regression model.
Residual Mean	Average value of the residuals.
Analysis of Variance (ANOVA)	Statical technique used to analyze the differences among group means in a sample.
Root Mean Squared Error (RMSE)	Measure of the difference between values predicted by a model or an estimator and the observed values.
Mean Absolute Error (MAE)	Measure of difference between two continuous variables, often used to assess the quality of a model's predictions.
Multicollinearity	Two or more predictor variables in a multiple regression model are highly correlated.
Variance Inflation Factor (VIF)	Index is used to measure the severity of multicollinearity in a regression analysis.

Table 2. Key terms used in the inferential Statistics section

Additional Analysis

To analyze the data further, plots 0 to 8 provide valuable insights into the dynamics of NFL games. Plot0, a boxplot of yards gained by week, shows varying team performances throughout the season, where weeks 19 to 22 display lower yards gained when compared to the rest. Plot 1, a scatter plot shows pass length and air yards of a scatter plot, suggesting that longer air yards are preferred for deep passes compared to short ones. Plot 2, which visualizes yards gained and yards after the catch, revealed that most yards gained are below 20 yards, and there is less likelihood of high-yard gains after catching the ball. Plot 3, a boxplot of passing yards by down, highlighted the challenges teams face in gaining yards during specific down situations, with first and third downs showing more significant yard gains when compared to the other downs. Plot 4 is a line plot of yards gained over time, suggesting potential shifts in team performance, particularly during later months of the year. Plot 5, a bar plot showcases yards gained by each home team with each down, underscoring the varying strategies adopted by different teams during different downs, with some teams showing preferences for certain play calls during downs situations. Plot 6 and 7 are stacked bar plots of play types by the home and away teams, emphasizing the differences in play strategies by teams at home and away games, with some teams showcasing distinct preferences for play types. Plot 8, a scatter plot with a linear regression line between passing_yards and air_yards, highlighted the correlation between these variables and potential trends in passing game strategies. Overall, the analysis sheds light on the strategic and performance aspects of NFL teams during games, showcasing variations and tendencies existing within the league.

Additional Time Series Analysis

The analysis of time series and ARIMA forecasting provided valuable insights to the NFL by revealing the overall trends in passing yards and yards gained. It highlights the potential fluctuations and uncertainties in these variables over time for strategic decision-making. The time series plot for yards gained demonstrates a steady progression with a noticeable shortening trend over months, showing potential for seasonal influences on offensive performance. The ARIMA forecast for yards gained shows the majority of data points are likely to cluster around the mean value of 4.034, with confidence intervals showing a range of fluctuations in yardage gained.

Furthermore, the ARIMA forecast plot illustrates the fluctuating nature of yards gamed over time, with occasional sharp declines in lost yards and intermittent peaks reaching up to 95 yards gained, providing teams with a comprehensive understanding of potential variability in offensive performance. Also, the time series plot for passing yards displays a relatively even distribution, except for a dip at -20, suggesting a stable passing performance over time. The ARIMA forecast for passing yards further supports the stability in passing yardage trends, showing the predictability of future passing performance. This data can aid teams in making informed decisions and optimizing their offensive strategies for enhanced competitiveness in the NFL.

Time Series	Sequence of data points, measured typically at successive points in time, determine trends of patterns.
Trends or Patterns	Direction in which data moves over time, often indicating upward, downward, or stable trends.
Forecast	Prediction of future values based on past and present data points.
Autoregressive Integrated Moving Average (ARIMA)	Statistical model used for analyzing and forecasting time series data, incorporating autoregressive, differencing, and moving average components.

Table 3. Key terms used in the additional time series analysis section

Discussion/Conclusion

The study of the NFL passing game has revealed crucial insights into the factors that affect passing yardage and the offensive performance’s temporal patterns. The multiple linear regression model has emphasized the significance of air yards and yards gained in shaping passing yards and underscores the importance of downfield passing and overall yardage gained in team strategies. However, the mode’s residual plots indicate potential deviations from the model assumptions, highlighting the need for further model refinement.

Additional analyses, including visualization and descriptive statistics, provide detailed insights into team performance, strategic tendencies, and play type preferences within the NFL. These insights could be pivotal for coaches, players, and management, enabling them to fine-tune their game plans, make data-driven decisions, and optimize offensive strategies for enhanced competitive performance. The implications of these findings could extend beyond the NFL, proving valuable lesions for other football leagues and potentially other sports, given the general applicability of strategic insights and performance analysis.

These results show importance, such as the relationships between variables. Air yards and yards gained influencing passing yards could be attributed to the strategic emphasis on deep passing and downfield plays in the NFL. Teams may prioritize gaining more yards through long passes, resulting in a positive correlation between air yards and passing yards. As such, the importance of yards gained could reflect the overall effectiveness of offensive strategies in advancing the ball downfield. Also, variable significance with certain variables, such as air_yards and yards_gained, might be due to their direct impact on overall offensive performance in NFL teams. These variables capture crucial aspects of passing plays, reflecting the interplay between the distance of the pass and the yards gained by the receiving player. Opposingly, the significance of variables might also be influenced by other external factors like game situations, defensive strategies, and specific player dynamics, which can affect the overall passing dynamics. It may also influenced by game situations, where certain variables are tied to specific game situations or play types. For example, highly influential during crucial down situations, such as third downs or red zone plays may exhibit different significant levels compared to other phases in the game, which can help understand the NFL passing game.

A surprising insight is how not the NFL has a consistently high passing yardage, as indicated by the mean and median passing yards of 10.93 and a standard deviation of 4.85, meaning that teams can execute successful passing plays, resulting in substantial yardage gains on average. Similarly, air yards and passing yards have coefficients of approximately 0.30 and 0.36, respectively, which demonstrates the role of these variables in determining total passing yards, emphasizing the significance of longer passes and overall yardage gained in the success of passing plays. The R-squared value of 0.5493 indicates that the model can predict 55% of the variability in passing yards, showing that the model is moderate to strongly predictive in explaining changes in passing yards based on air yards and yards gained. Furthermore, the cross-validated results confirm the mode’s reliability and accuracy in predicting passing yards, with a root mean squared error (RMSE) of 3.256485 and a mean absolute error (MAE) of 2.248118, based on the input variables of air yards and yards gained. Additional analysis of various plots reveals several intriguing findings, including the variation in team performances during different weeks, the preference for deep passes over short ones, and the likelihood of gaining yards after catching the ball. The correlation between air_yards and passing yards highlights the range of air yards and their distribution pattern, providing valuable insights into passing game strategies. The ARIMA forecasting results demonstrate the consistent and predictable nature of passing performance in the NFL while also indicating the potential impacts of external factors through sharp declines and peaks, emphasizing the need for adaptable strategies to account for potential variability. Overall, the results provide valuable insights into the dynamics and intricacies of passing plays in the NFL, offering implications for strategic decision-making and further research in the field.

It is important to note the limitations of this study, as it only focuses on passing yards and doesn’t consider the full complexity of the game. Future research should broaden the analysis to include a range of performance metrics, incorporate defensive strategies and player-specific analysis, and use advanced statistical methods to capture the intricacies of team dynamics. Additionally, exploring the influence of external factors such as weather conditions, stadium types, luck, and game situations could provide further context for understanding the nuances of the NFL passing game. The time capacity of the assignment was affected by the need to pick a model to combat the residual model's performance, which could be improved in the future. Collecting additional data sources and conducting more granular analyses could also strengthen the robustness of the findings and provide a more holistic understanding of the NFL passing game. Furthermore, broadening the analysis to include defensive strategies and player-specific dynamics would offer a more comprehensive perspective on the multifaceted nature of team performance in the NFL. Incorporating advanced statistical techniques, such as machine learning algorithms or causal inference methods, could provide deeper insights into the causal relationships and potential driving forces behind passing game dynamics.

References

Bevans, R. (2023, June 22). *Multiple Linear Regression | A Quick Guide (Examples)*. Scribbr. https://www.scribbr.com/statistics/multiple-linear-regression/

Joshi, A. V. (2022). Time series models. In *Springer eBooks* (pp. 139–148). https://doi.org/10.1007/978-3-031-12282-8_12

Shook, N. (2023, June 1). Next Gen Stats' top 10 NFL deep passers of 2022: Geno Smith, Tua Tagovailoa excel at airing it out. *NFL.com*. https://www.nfl.com/news/next-gen-stats-top-10-nfl-deep-passers-of-2022-geno-smith-tua-tagovailoa-excel-a

Brickwallblitz. (2021, February 26). *The 2020-21 Deep Ball Project (Part 1/3)*. Brick Wall Blitz. https://brickwallblitz.com/2021/02/16/the-2020-21-deep-ball-project-part-1-3/

Meyer, J. (2022, April 2). Multiple regression analysis for predicting the 40 yard dash for NFL prospects. *Medium*. https://medium.com/@jdmeyer05/multiple-regression-analysis-for-predicting-the-40-yard-dash-for-nfl-prospects-9d92d9a6f58b

Zach. (2020, November 16). *Introduction to multiple linear regression*. Statology. https://www.statology.org/multiple-linear-regression/

Zach. (2022, June 9). *How to perform multiple linear regression in R*. Statology. https://www.statology.org/multiple-linear-regression-r/

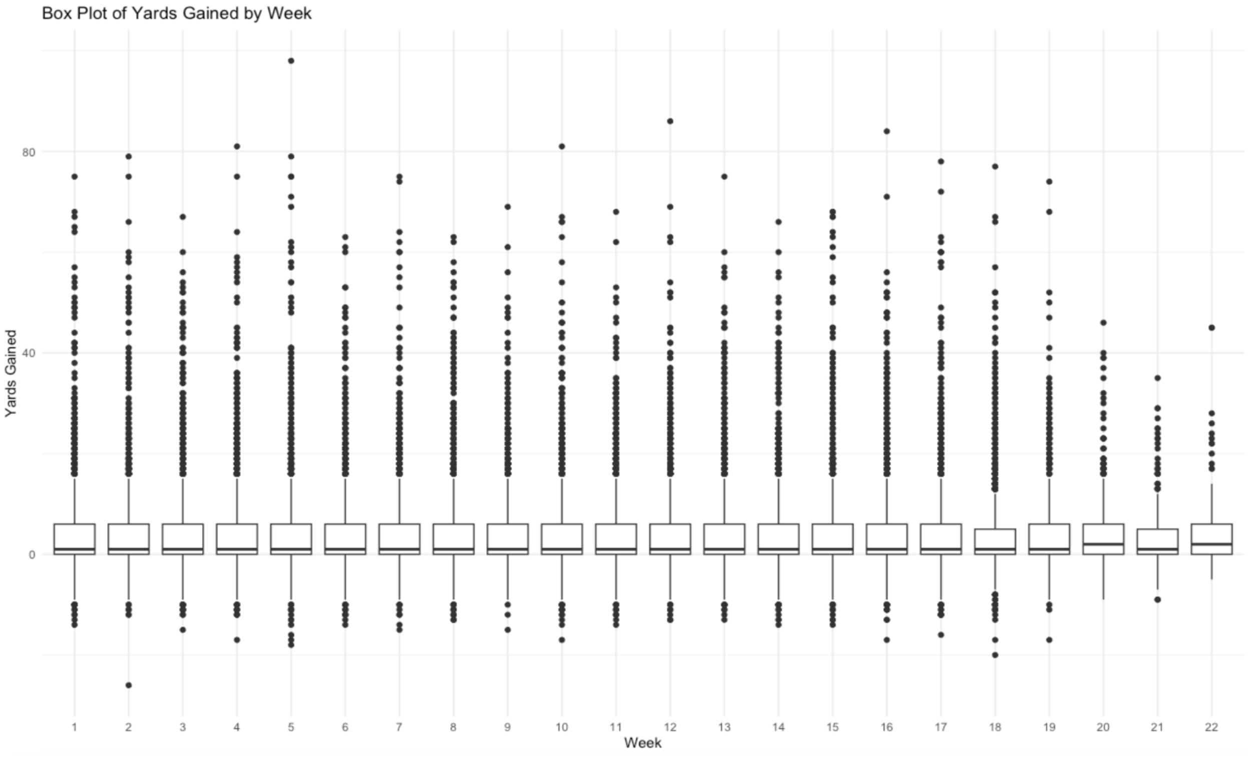

Figure 1. Boxplot showing weekly yards gained variation during the NFL season

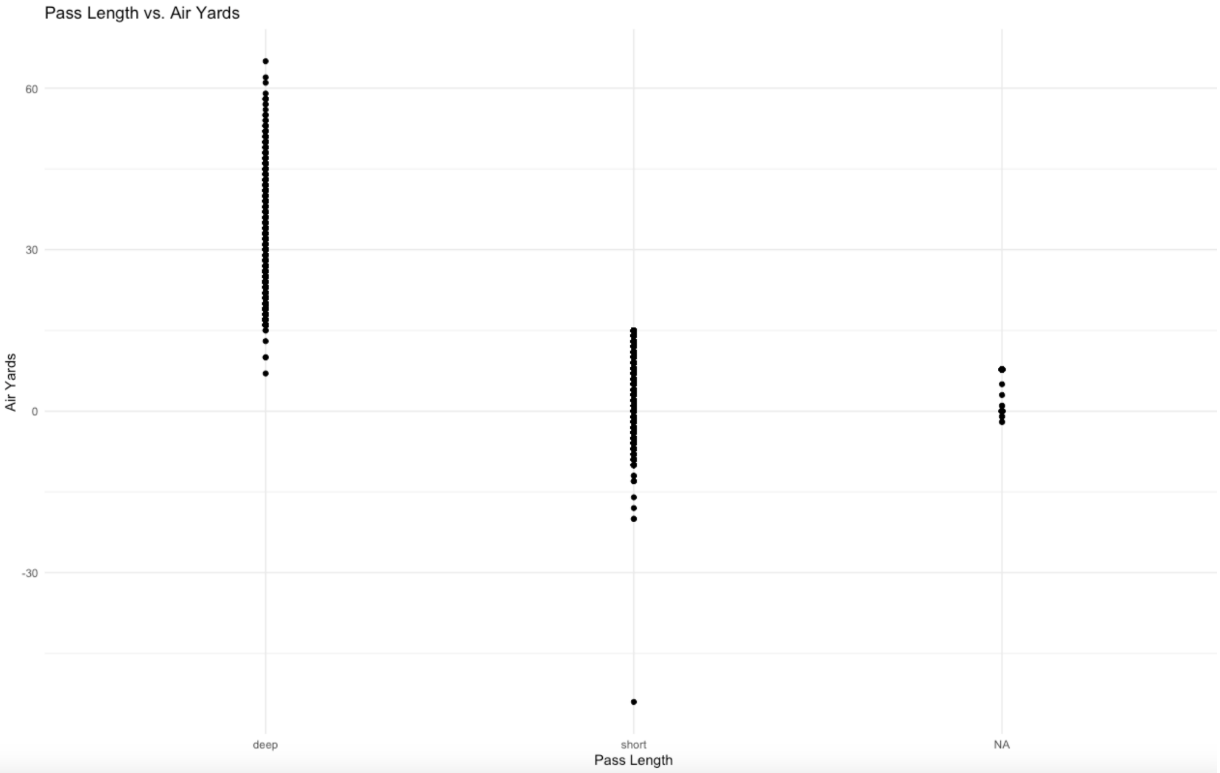

Figure 2. Scatter plot demonstrating the relationship between pass length and air yard

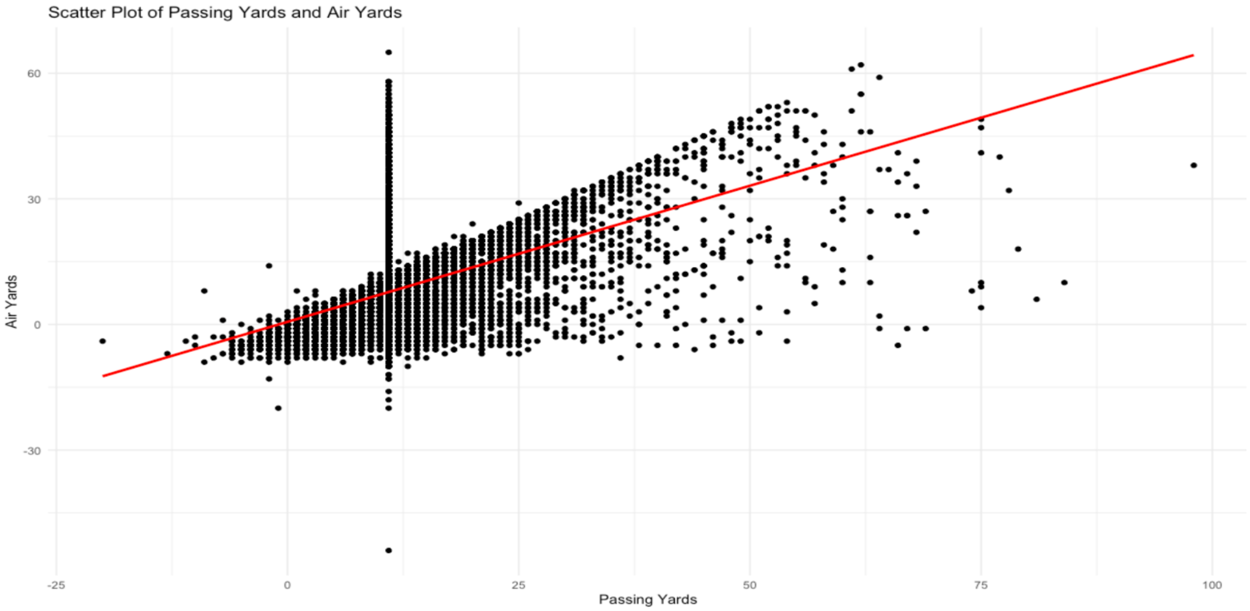

Figure 3. Line plot illustrating the progression of yards gained over time during the NFL season