Breaking Down the NFL: How Offensive Yards Shape Game Outcomes*

Examining the Trends in Yardage of Winning NFL Teams Using Logistic Regression

Alexander Guarasci

April 3, 2024

This study examines the strategy behind National Football League (NFL) games, focusing on how teams' offensive yardage—gained through both rushing and passing—affects their chances of winning using two logistic regression models. Analyzing play-by-play data from the 2023 NFL season, we discovered that while both rushing and passing yards contribute to game victories, a stronger rushing attack significantly increases a team's likelihood of winning. These findings highlight the critical role of offensive strategy in football, providing valuable insights for teams looking to optimize their game plans for better outcomes.

Introduction

When it comes to professional football, there has long been conjecture that a team's performance in games is closely correlated with their offensive yardage total, which includes both passing and running yards. In order to find patterns and insights that could affect team strategy and fan comprehension of the game dynamics, this study conducts a thorough statistical analysis of the complex link between various offensive yard measurements and the probability of winning NFL games. This research attempts to quantify the effect of offensive performance on game outcomes by methodically analyzing play-by-play data for the 2023 NFL season. It provides a thorough summary of how offensive yardage affects a team's victory both throughout the entire game and when focusing only on the statistics for the fourth quarter. Ultimately trying to answer the questions, how do offensive yards impact the outcome of an NFL game, and what is more valuable passing yards or rushing yards[1]?

*Code and data underpinning this paper are available at: https://github.com/AlexanderG123/nfl_analysis

[1]Rushing yards and running yards are used interchangeably throughout this paper

The constant discussion among coaches, commentators, and fans over the best tactics for winning football games is what inspired this research. Although conventional wisdom has frequently emphasized the significance of a balanced offence—one that features a dynamic rushing attack in addition to a solid passing game—the empirical evidence that backs up this claim has been largely anecdotal. This paper presents a data-driven investigation of the correlations between various offensive methods and winning games. This analysis will help to provide insights into tactical choices and advance our knowledge of football analytics.

This analysis was conducted through the R programming language (R Core Team 2022) and supplemented with the analytical tools Tidyverse (Wickham et al. 2019) and caret (Kuhn and Max 2008). The dataset was assembled from play-by-play records of NFL games from NFLVerse (Carl et al. 2023). The data is analyzed using statistical methods, such as logistic regression, which provides insights into the relative value of passing and rushing in securing game wins. This method opens the door for further study in sports analytics and offensive strategy optimization while also illuminating the prevailing dynamics of effective offensive methods.

This study aims to make a substantial contribution to the conversation about sports analytics by offering a strong analytical framework for examining the connection between NFL game-winning and yards. It reveals insights into offensive strategy through in-depth statistical analysis and model-building, providing information for analysts, players and fans. The results of this study show how important data analysis is to sports strategy and how quantitative methods can potentially help improve competitive advantages in professional football.

2 Data

The data used in this paper was taken from nflverse (Carl et al. 2023), which consists of "a set of packages dedicated to data of the National Football League."

2.1 The Dataset

The foundation of the dataset that was ultimately used in this paper is the play-by-play data from across the 2023 NFL season. This includes an unbelievably dense array of information from every snap over the entirety of the season consisting of 372 variables. These include: seconds left in the half, touchdown probability, expected points added, the probability of taking a safety, the position the snap was taken in, a description of the play, and many, many more. In order to filter this information the play-by-play data was cleaned and new variables were created.

Other statistics that were taken into consideration but ultimately not used included more detailed play-by-play assessments and analytics on the performance of individual players. It was

determined that the comprehensive dataset with an emphasis on team-level performance indicators was the most suitable for investigating the overall correlation between offensive yardage and game results. This method avoids player-specific variability and enables a macro-level analysis, which is in accordance with the study's goal of outlining general strategic insights.

2.2 Variables of interest

This comprehensive dataset encapsulates a wide array of variables, central to which are "**RushingYards**", "**PassingYards**", "**TotalYards**", "**home_score**", "**away_score**", "**Winner**", "**Win**", "**Q4_RushingYards**", "**Q4_PassingYards**" and "**Q4_TotalYards**". **RushingYards** and **PassingYards** represent the yards gained by a team through rushing and passing plays, respectively, within a single game, while **Q4_RushingYards** and **Q4_PassingYards** represent the rushing and passing yards in the fourth-quarter. **RushingYards** is taken as the sum of the variable "**rushing_yards**" from the original dataset with the same "**game_id**" (so that we get the rushing yards from each game) and "**posteam**" (so each team that plays gets has their rushing yards tracked). **PassingYards**, **Q4_RushingYards** and **Q4_PassingYards** are similarly constructed. **TotalYards** is a constructed variable, representing the sum of **RushingYards** and **PassingYards**, devised to encapsulate a team's overall offensive performance. **Q4_TotalYards** is constructed as the sum of **Q4_RushingYards** and **Q4_PassingYards**. The scores of the home and away teams are captured by **home_score** and **away_score**, respectively, which is data that is provided and that does not need to be constructed. The **Winner** is determined based on the comparison of these scores, classified as the team that won (for example, "Bal" if the Baltimore Ravens won). Finally, the binary **Win** variable signifies the outcome from the perspective of the team being analyzed (1 for a win, 0 for a loss, and "TIE" for a tie, but because there were no ties last season there are no instances of "TIE").

Summary statistics were generated to provide an overarching view of the data's distribution and central tendencies. These statistics reveal an average of approximately 113 rushing yards and 238 passing yards per game, with total yards averaging around 350. The distribution of these variables, along with the game outcomes (Win), was visually examined through histograms and box plots, elucidating the central tendencies and variability within the data. These visual explorations further affirmed the preliminary assumption that a higher aggregate of offensive yards tends to correlate with a higher likelihood of winning.

The measurement aspect of this data is something that can be scrutinized. If the reader of this paper has ever watched an NFL game, they are likely very aware of the lack of certainty with regard to where the football is placed. A player gains yards until they are tackled. Following the tackle the player hands the ball to the referee who then places it where he or she believes the ball was when the player went down, and which is where the next play begins. This spot is where the data is recorded. Although this is a severely flawed technique that often infuriates viewers, it is not something that concerns our analysis because we are curious about the end

results, how each gained yard benefits a team. We do not care about this difference because it does not impact the results of a game. For example, if a player gets six yards on a play but the referee only spots him three yards, the end result is a three-yard gain, the six yards will have no impact on the outcome.

3 Models

Two logistic regression models were used in our study to analyze the NFL game dynamics, with a particular focus on the relationship between a team's cumulative running and passing yards and its likelihood of winning. Both the full game and just the fourth quarter are catered to by these models, which provide a multi-layered understanding of strategic applications at different points of the game.

3.1 Model 1: Overall Game Analysis

The first logistic regression model scrutinizes the entire game, using rushing and passing yards as predictor variables to estimate the odds of winning. This first model, Equation 1, expressed as

$$\log\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \beta_1 \cdot \text{RushingYards} + \beta_2 \cdot \text{PassingYards} \tag{1}$$

, where p1 denotes the probability of winning, facilitates a broad analysis of strategic effectiveness. The coefficients, representing rushing and passing yards respectively, illuminate the value of each yard gained in the context of winning. The positive coefficient for rushing yards implies a direct correlation between a stronger rushing offence and the likelihood of victory, underscoring the strategic advantage of a balanced offensive game. This model is apt for providing a holistic view of game strategy, allowing analysts and teams to gauge the overall impact of their offensive efforts throughout the match.

3.2 Model 2: Fourth Quarter Focus

The second logistic regression model, Equation 2, narrows the focus to the fourth quarter, a critical period where games are often decided. By analyzing the same predictors but confined to the final quarter, this model,

$$\log\left(\frac{p_2}{1-p_2}\right) = \beta_3 + \beta_4 \cdot \text{Q4\_RushingYards} + \beta_5 \cdot \text{Q4\_PassingYards} \tag{2}$$

, offers insights into the tactical shifts that dominate the endgame. Unlike the overall game model, the fourth-quarter model might reveal a nuanced strategic landscape, such as a negative coefficient for passing yards, indicating the potential pitfalls of a pass-heavy approach

under pressure. This focused analysis is instrumental in understanding the dynamics of clutch gameplay, shedding light on the adjustments teams make in response to the scoreboard and the clock.

3.3 Justification and Appropriateness

The choice of logistic regression is prudent in light of the binary structure of the outcome variable (win or loss) and the models' ability to convey the connection between playing strategies (rushing and passing) and winning odds. The interpretability of logistic regression emphasizes its suitability even more; the exponential of the coefficients accurately measures the change in odds that come with every extra yard gained, whether by passing or rushing. This gives you a concise, useful grasp of how various aspects of the game affect your chances of winning.

Furthermore, the choice to use two different models takes into account the intricate, multidimensional character of football strategy. The fourth-quarter model focuses on the crucial choices and plays that frequently determine the result, while the overall game model reflects the persistent efforts and strategies that mould the game's early and middle phases. This dual-model method enables a thorough analysis that acknowledges the urgency of the game's closing moments as well as the marathon nature of establishing a lead.

In summary, the use of logistic regression models to analyze NFL game outcomes offers a robust methodological framework for dissecting the strategic components of football. By evaluating the entire game and the pivotal fourth quarter separately, these models provide nuanced insights that can guide teams in refining their strategies for sustained success and clutch performance.

4 Results:

As an overview of the whole data set, of the 285 games played in the NFL regular season and playoffs, there were 90 games where a team rushed for fewer yards than their opponent and won while there were 109 games where a team threw for fewer yards than their opponent and won. Prioritizing the rushing attack is further bolstered by the model which suggests that a rushing yard is more valuable than a passing yard. In only 72 games a team rushed and passed for more yards than their opponent and won the game and in 14 games a team rushed for and passed for fewer yards while winning the game.

4.1 Graphical Results

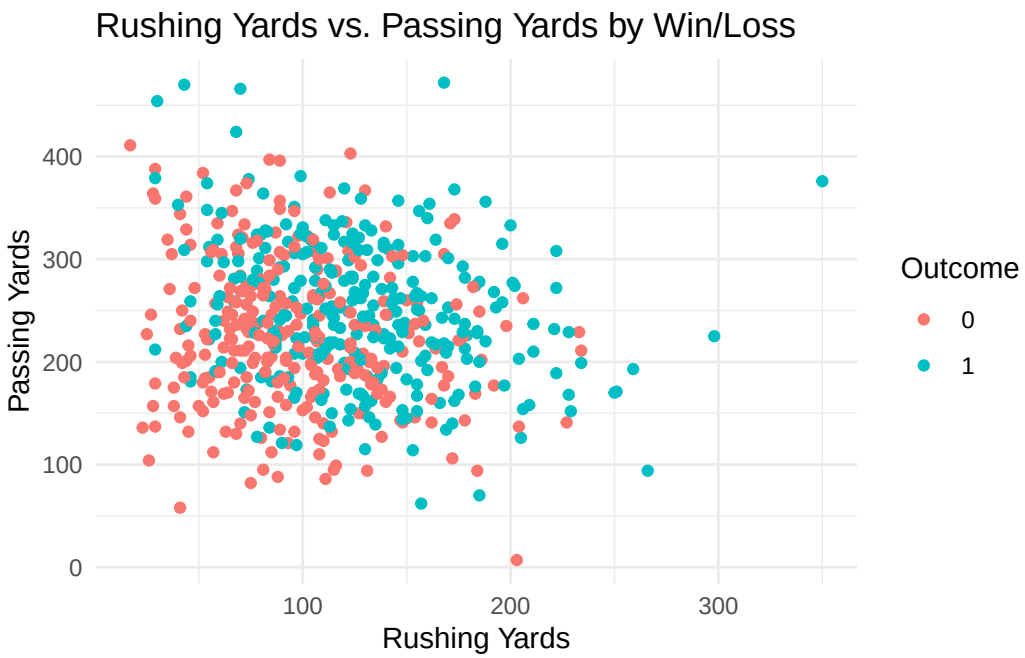

Figure 1: The Rushing and Passing Yards of Every Game in 2023

Figure 1 shows the rushing yards and passing yards of winning teams over the 2023 NFL season. In examining the graph it is clear that the overall trend, more often than not, is that more rushing yards and more passing yards increase the likelihood of winning. The biggest outlier game, which resulted in the Miami Dolphins Rushing for 350 yards and passing for 376 yards on the Denver Broncos, led to a 70 to 20 victory for the Dolphins. Prior to that game, no team in the 21st century had scored 70 points in a game, this anecdotally shows how being able to run and pass the ball effectively is exceedingly helpful when it comes to scoring points and ultimately winning ball games.

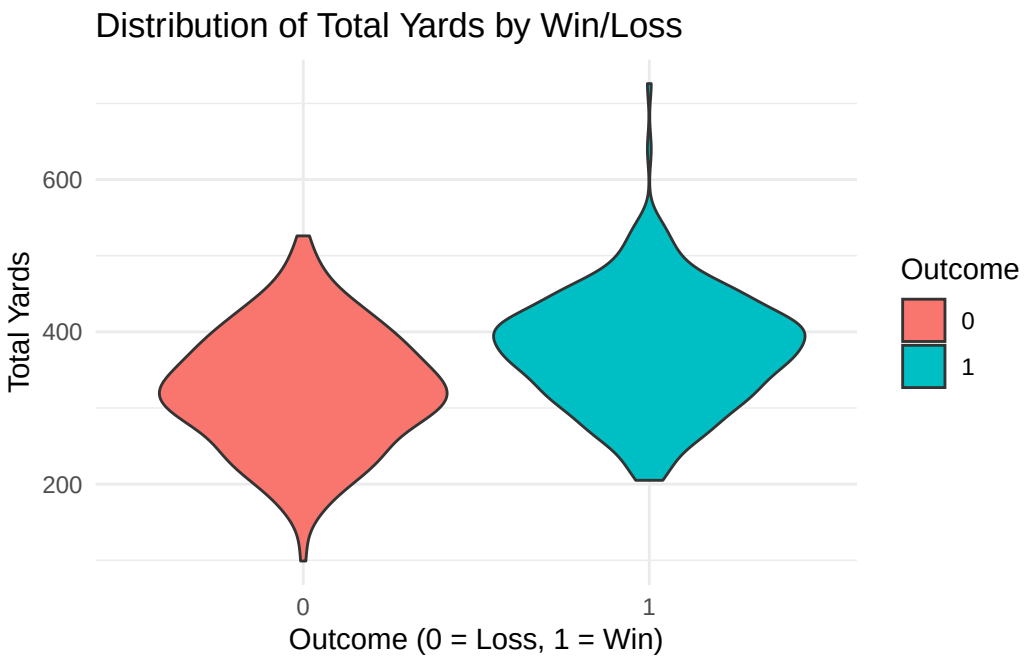

Figure 2: Violin Plot of Total Yards by Wins and Losses

Figure 2 supports this idea, it is a violin plot of the total yards of winning and losing teams (total yards is the sum of passing yards and rushing yards). What it shows is that winning teams never had fewer than 200 total yards, and had more total yardage on average than the teams that lost. It also shows that most winning teams had around 400 total yards while losing teams had just over 350 yards.

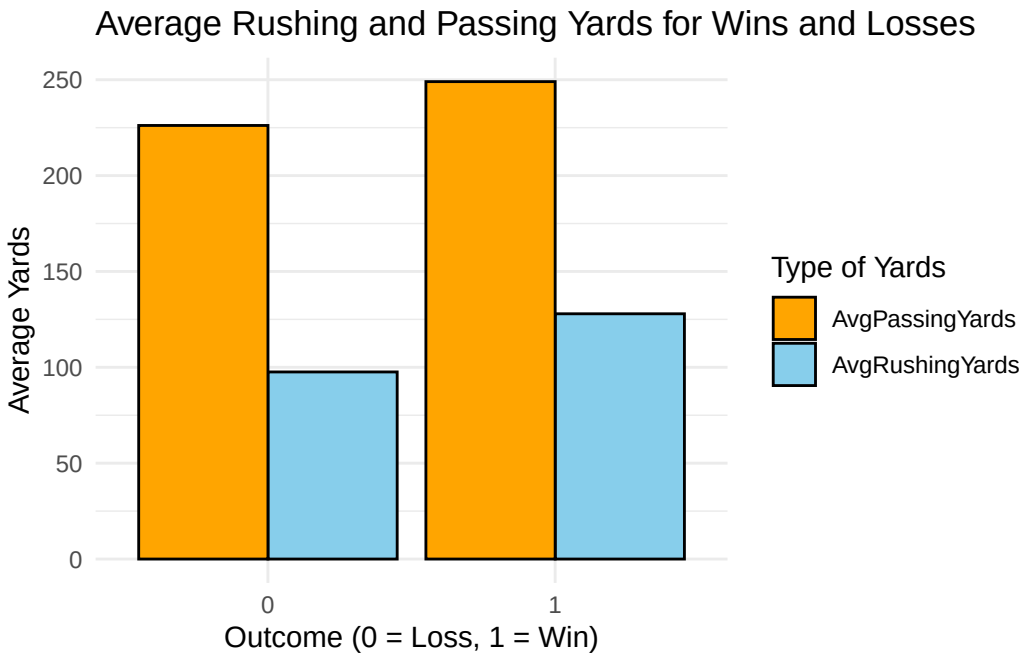

Figure 3: Average Yards by Wins and Losses

This is further reinforced by Figure 3 which shows the average passing and rushing yards of winning and losing teams. As you can see, the winning teams rushed and passed for more yards. The results become more interesting when we break them down in the fourth quarter.

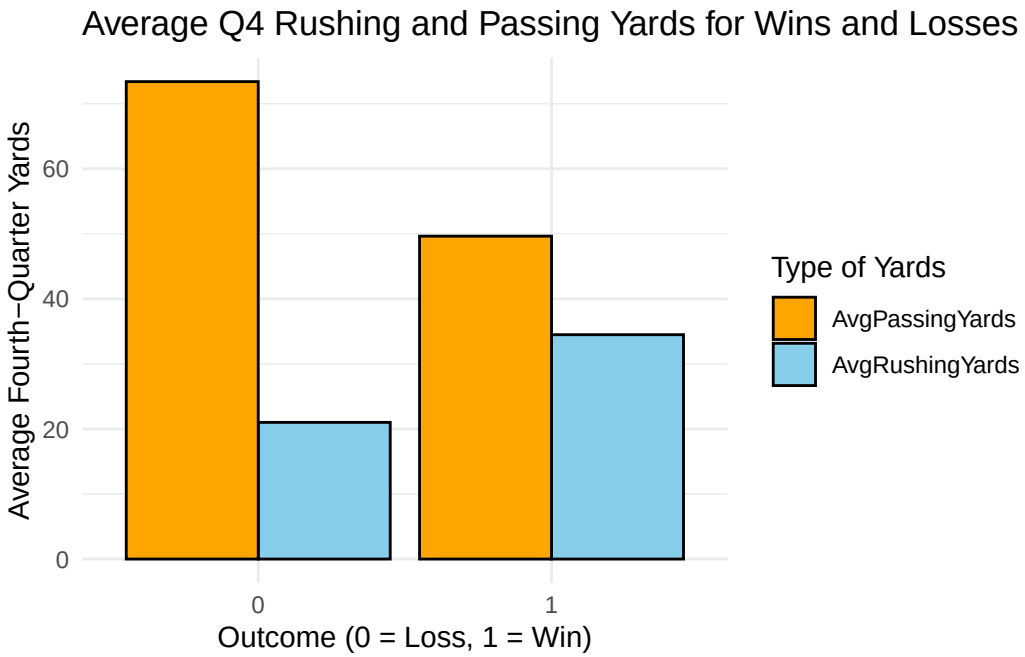

Figure 4: Average Fourth-Quarter Yards by Wins and Losses

Figure 4 shows a stark difference in pass yards and rushing yards of the losing team. This trend is substantially more pronounced than Figure 3 is and provides a much more interesting backdrop, teams that lose football games pass the ball substantially more than the teams that win, and in particular they do so in the fourth quarter.

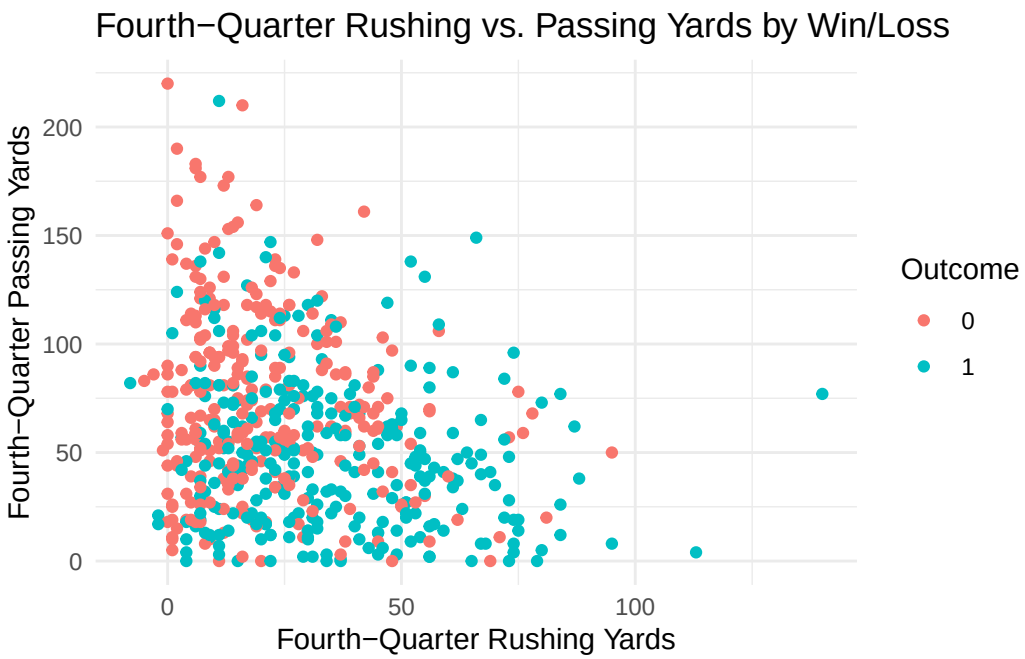

Figure 5: The Fourth-Quarter Rushing and Passing Yards of Every Game in 2023

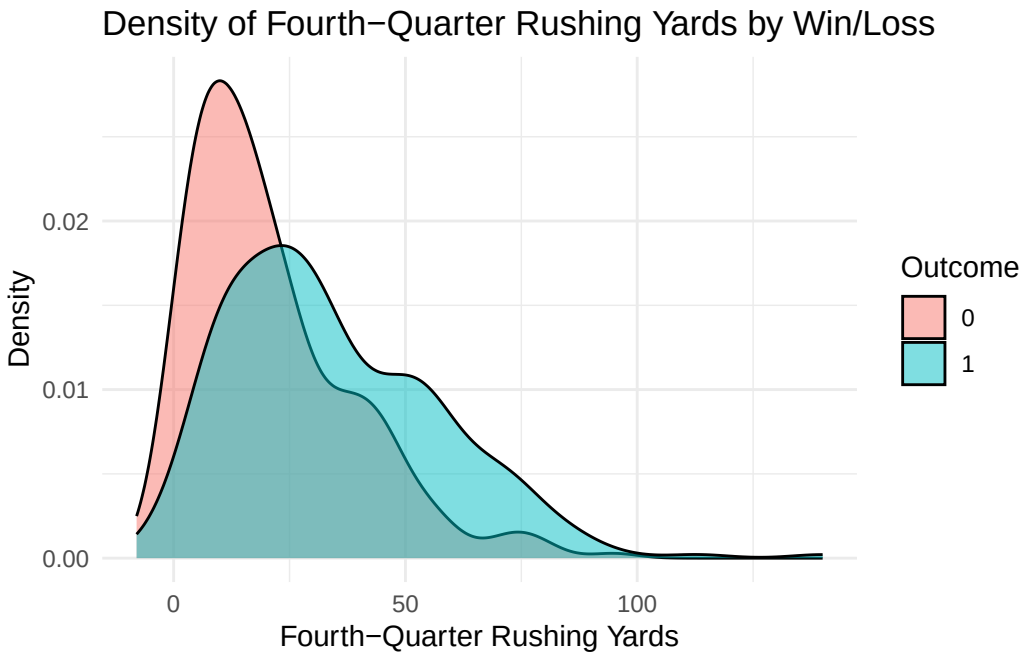

Figure 6: Distribution of Fourth-Quarter Rushing Yards by Wins and Losses

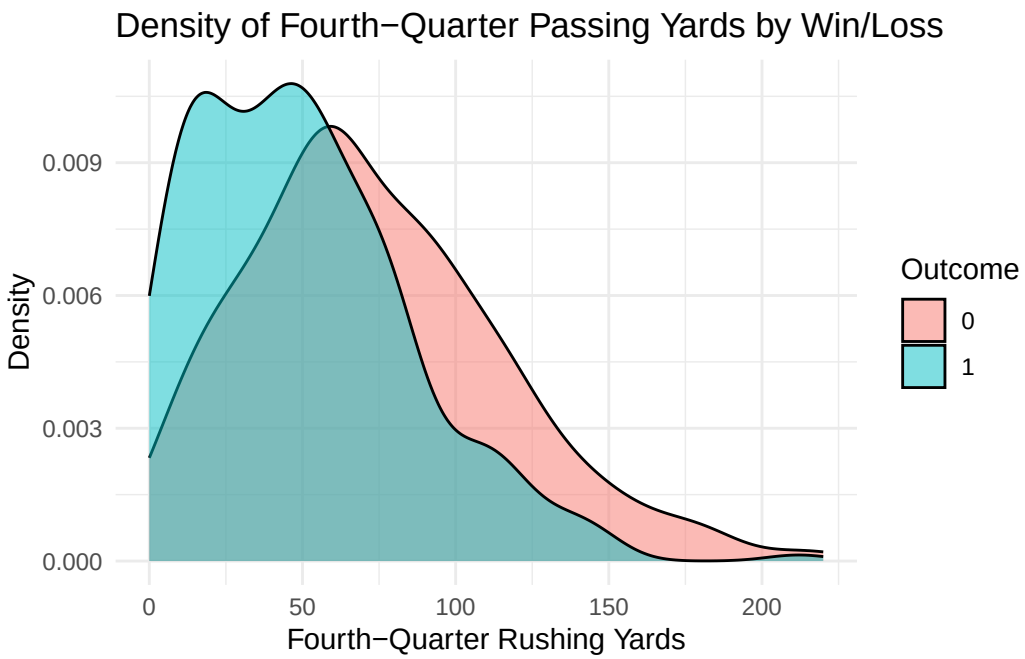

Figure 7: The Distribution of Fourth-Quarter Passing Yards by Wins and Losses

Comparing Figure 5 with Figure 1, there is quite a contrast. Whereas Figure 1 showed that winning teams ran and threw the ball more, Figure 5 shows the winning teams hugging the x-axis (more fourth-quarter rushing yards and fourth-quarter fewer passing yards), with losing teams hugging the y-axis (fewer fourth-quarter rushing yards and more fourth-quarter passing yards). Figure 6 and Figure 7 show the distributions of fourth-quarter rushing and passing yards by winning and losing teams and both show how pronounced this trend is.

4.2 Model Results

This paper relied on two models. The first was a logistic regression of overall game data while the second was a logistic regression on purely fourth quarter data.

The regression on overall data indicated a positive relationship between rushing yards and passing yards and winning the game, with rushing yards being considered more valuable with regard to winning. This model was able to correctly predict 65.79% of outcomes in the test data, predicting 71.93% of actual wins and 59.65% of actual losses. It demonstrated a meaningful ability to predict game outcomes based on rushing and passing yards but there was room for improvement.

The second regression, using only fourth-quarter data, had more interesting results. Similar to the first model, there was a positive relationship between rushing yards and winning, but with fourth-quarter rushing yards being more valuable than overall game rushing yards. Interestingly, there was a negative relationship between passing yardage and winning the game, in other words, the more passing yards a team had in the fourth quarter, the more likely that team was to lose the game. This model was more accurately able to predict the outcome of games, correctly predicting 78.95% of winners and 66.67% of losers, for an overall balanced accuracy of 72.81%.

5 Discussion

The analysis and models presented in this paper provide a deep dive into the strategic dynamics of NFL games, with a particular focus on the impact of rushing and passing strategies on game outcomes. Through the lens of logistic regression models—both encompassing overall game performance and isolating the critical fourth quarter—we uncover nuanced insights into the traditional adage of football: "the run game wins matches." The findings affirm this notion but also highlight the complexity behind strategic decisions made on the field, especially under the pressure of the game's closing moments.

One significant insight this paper brings to light is the tangible value of a balanced attack in securing victories. The overall game model suggests a positive correlation between both rushing and passing yards with winning, reinforcing the importance of versatility in offense. This is a testament to the complexity of football strategy, where being unpredictable and capable in both facets of the game complicates the defense's task. However, the differential valuation of rushing over passing yards in the models indicates a slight strategic edge for the ground game, potentially due to the clock management and lower turnover risks associated with rushing.

The fourth-quarter model, however, opens a window into the high-stakes decision-making that defines tight contests. The negative relationship between fourth-quarter passing yards and winning underscores a common game scenario: teams behind on the scoreboard resort to the air to catch up quickly. While this strategy is rational and sometimes the only option, it is less likely to result in a win, as indicated by the model. This outcome may also reflect the defensive adjustments expecting passes, further decreasing the efficiency of a one-dimensional approach.

The study is not without its limitations. The reliance on aggregate statistics like total yards overlooks the game's situational context, such as the down-and-distance, the impact of turnovers, and special teams' play, which can significantly influence game outcomes. Moreover, the models do not account for the intricacies of clock management, a critical aspect of late-game strategy. Future research could benefit from incorporating these elements into more sophisticated models, perhaps employing play-by-play data to capture the game's flow and strategic shifts more accurately.

Looking forward, the exploration of strategic dynamics in NFL games could be enriched by integrating advanced metrics such as Expected Points Added (EPA) and Win Probability Added (WPA), which provide a more nuanced understanding of each play's impact. Additionally, qualitative analyses of coaching decisions, player performances in clutch situations, and the psychological aspects of late-game pressure could offer comprehensive insights into the art and science of winning football games. As the sport continues to evolve, so too will the strategies that define its outcomes, necessitating ongoing analysis and adaptation by teams and analysts alike.

References

Carl, Sebastian, Ben Baldwin, Lee Sharpe, Tan Ho, and John Edwards. 2023. *Nflverse: Easily Install and Load the 'Nflverse'*. https://CRAN.R-project.org/package=nflverse.

Kuhn, and Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.