Predicting NFL Team Success Based on Fantasy Football Player Performance: Exploratory Data Analysis

Problem Statement:

"Can Key Fantasy Football Performance Metrics Predict NFL Team Success? Analyzing the Impact of Yards Gained from Scrimmage in Determining Team Win-Loss Records"

Expanded Description:

In the ever-evolving world of sports analytics, fantasy football has emerged as a powerful tool for fans and analysts alike to engage with the NFL on a deeper level. Fantasy football scores are built on a foundation of real-world player statistics, reflecting key performance metrics such as yards gained, touchdowns scored, and turnovers. These metrics offer a direct glimpse into a player's contribution on the field and, by extension, their potential impact on their team's success.

However, while fantasy football focuses on individual statistics, the overall success of an NFL team is influenced by a complex interplay of factors, including team strategy, coaching decisions, and the performance of the defense and special teams. This project seeks to bridge the gap between individual player performance in fantasy football and the collective outcomes of NFL teams by asking a crucial question: **Can we predict an NFL team's win-loss record based on the fantasy football performance of its key players, analyzed on a position-by-position basis?**

"Yards Gained," in particular, is a statistic that often goes overlooked. Everyone knows that touchdowns put points on the board and win games, but "Yards Gained" is what keeps the ball moving and the offense on the field. This metric is the lifeblood of a successful offense, as it directly correlates with a team's ability to control the clock, sustain drives, and maintain momentum throughout a game. By analyzing how "Yards Gained" by key positions—such as quarterbacks, running backs, and wide receivers—impacts the overall team performance, we can gain valuable insights into the hidden drivers of NFL success, potentially revealing a predictive relationship between these individual contributions and a team's win-loss record.

To explore this question, the project will leverage publicly available NFL and fantasy football data, focusing on recent seasons and key positions such as quarterback (QB), running back (RB), wide receiver (WR), and tight end (TE). By conducting a position-by-position analysis, the goal is to identify whether “Yards Gained” through passing, rushing, or receiving is the most significant predictors of a team's success in terms of their win-loss record.

The findings from this analysis will have far-reaching implications for both fantasy football enthusiasts and NFL teams. For fantasy football managers, understanding which positions and metrics are most closely tied to team success could inform draft strategies and player trades, offering a competitive edge in their leagues. For NFL

teams, the insights could highlight the critical importance of certain player roles and metrics in achieving a winning season, guiding player development and game strategy decisions.

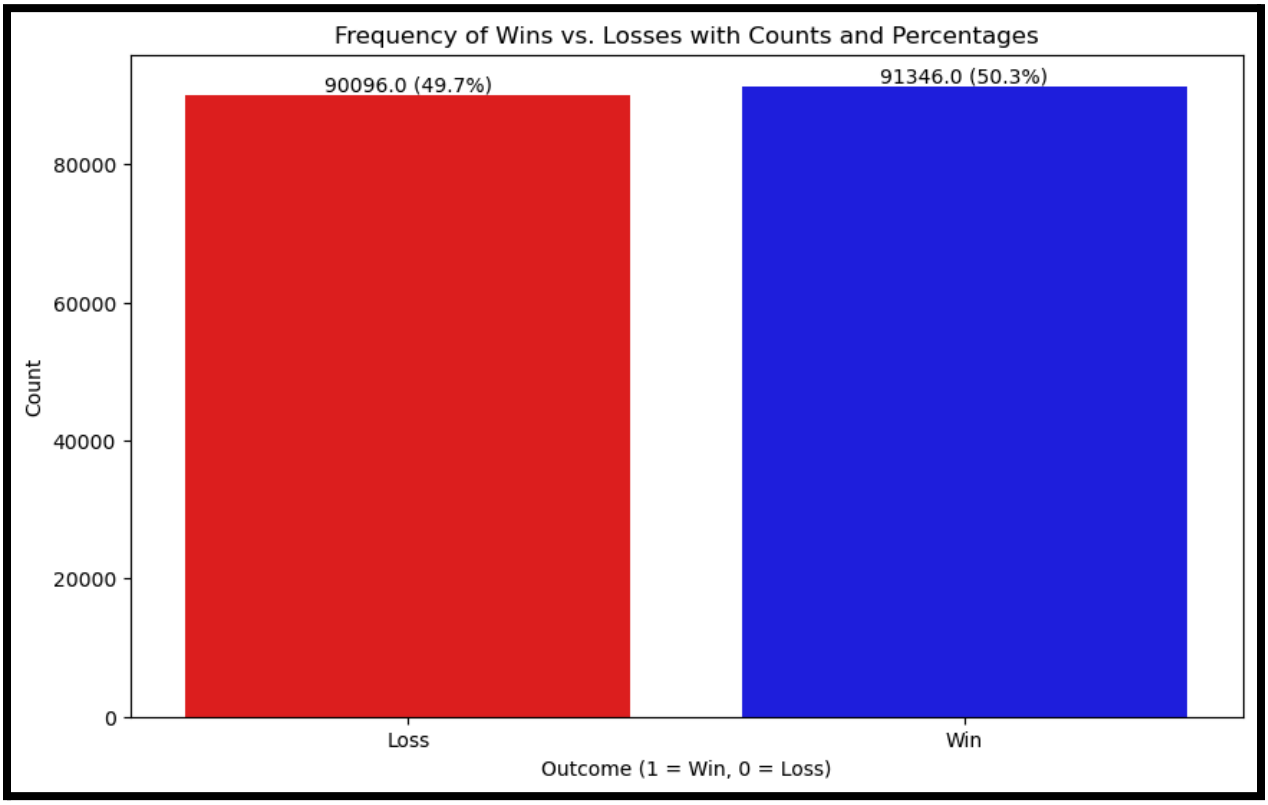
In conclusion, this project offers a unique opportunity to understand the dynamics of NFL team success through the lens of fantasy football. By analyzing the intersection of individual player performance and team outcomes, we aim to provide actionable insights that enhance decision-making in both fantasy sports and professional football. Whether you're a fantasy football manager seeking to optimize your team or an NFL analyst striving to understand the factors behind a winning season, this project will offer valuable perspectives on how individual efforts contribute to collective success.

Exploratory Data Analysis

Is the Data Valid?

The univariate chart of the target (wins and losses) (Figure 1) shows that the data is acceptable and valid. The target variable is normally distributed and does not present data imbalance. The target variable is approximately split evenly between wins and losses.

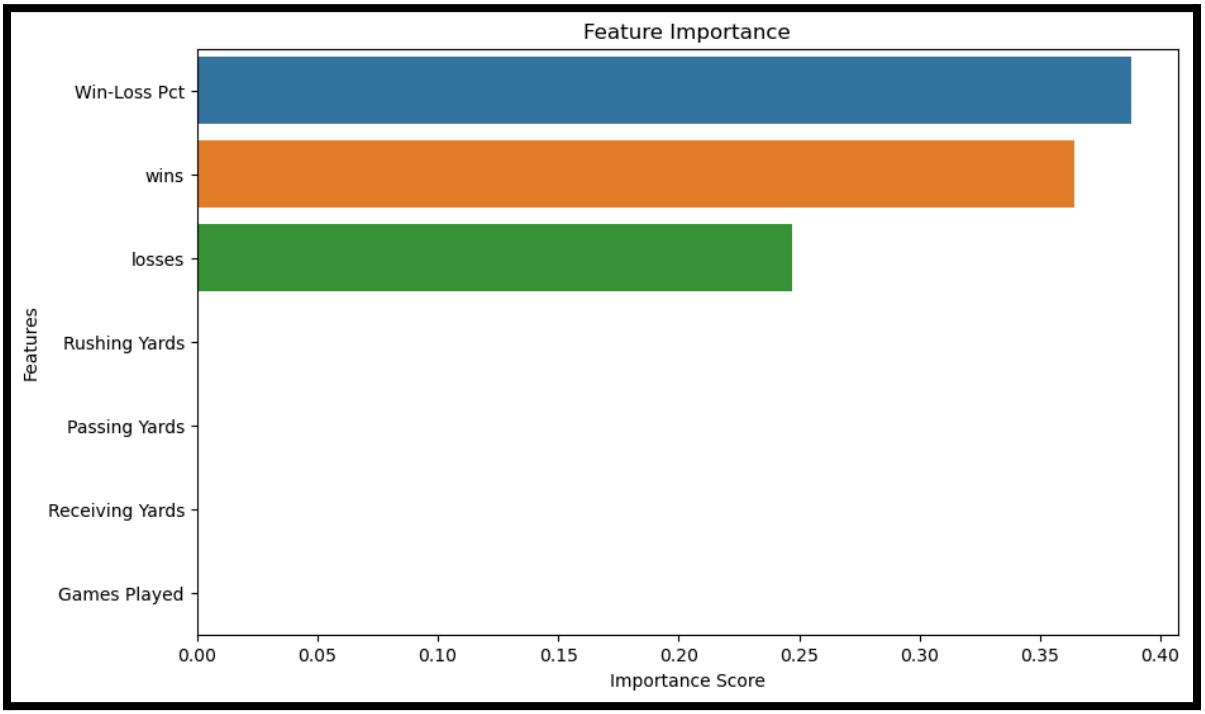
Figure 1: Frequency of Wins and Losses



Which Variables Impact Wins and Losses?

One of the key advantages of a Random Forest is its ability to rank the importance of different features. The model tells you which features (variables) are most important in making predictions. In this case, the model should show how important passing yards, rushing yards, and receiving yards are in predicting the game outcome. The model shows that there is very little relationship between winning and yards gained from scrimmage.

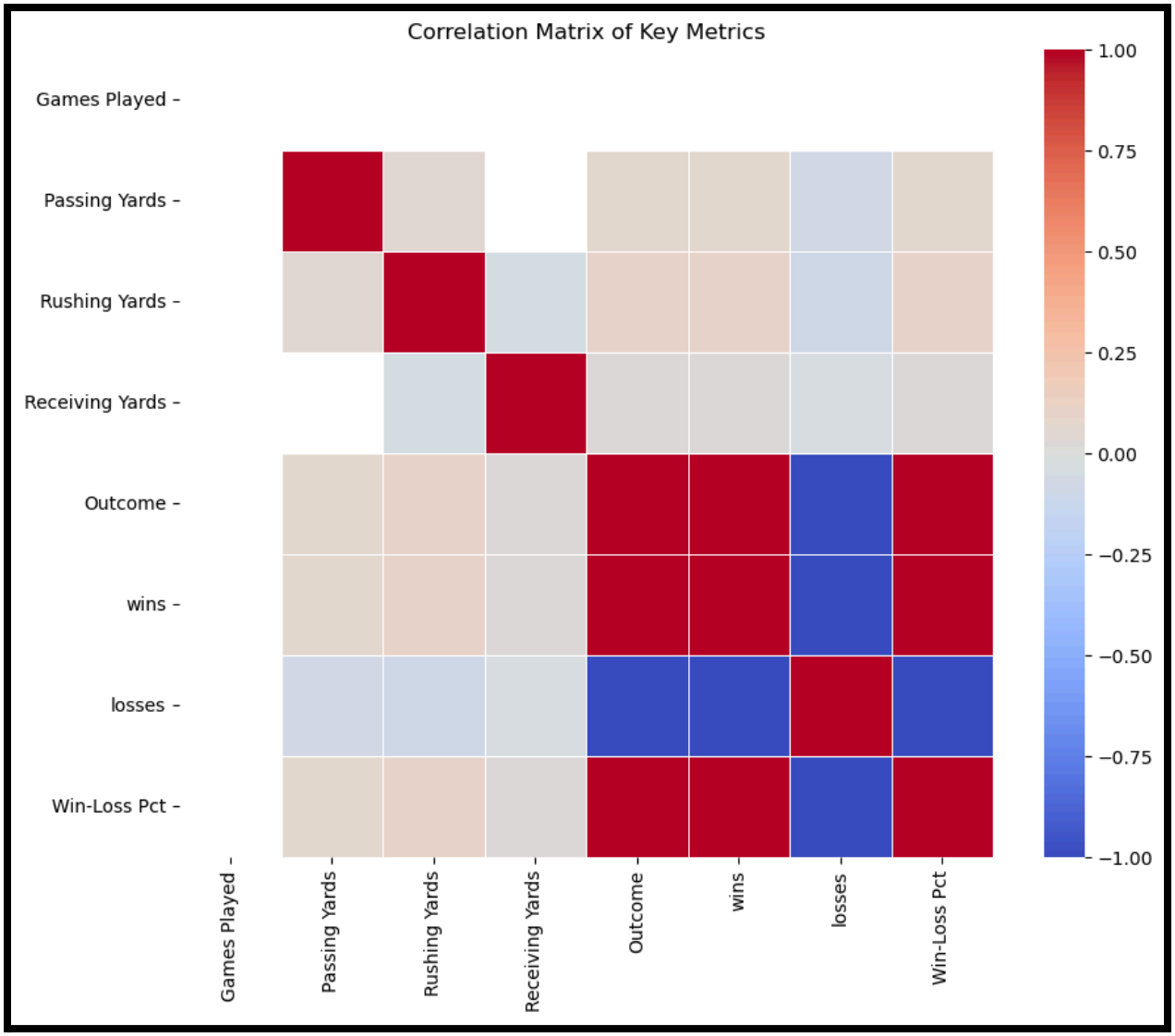
Figure 2: Random Forest Generator



Is there Correlation in the Data?

The correlation matrix (heatmap) also shows no correlation between yards gained, whether through rushing, passing or receiving, and wins/losses (a value around 0 suggests no linear relationship between the variables). Interestingly, there is a negative correlation (when values on the heatmap are closer to -1.0) between passing and rushing yards gained and losses. This tells us that as passing or rushing yards increase, there is a higher probability of losing the game.

Figure 3: Heat Map

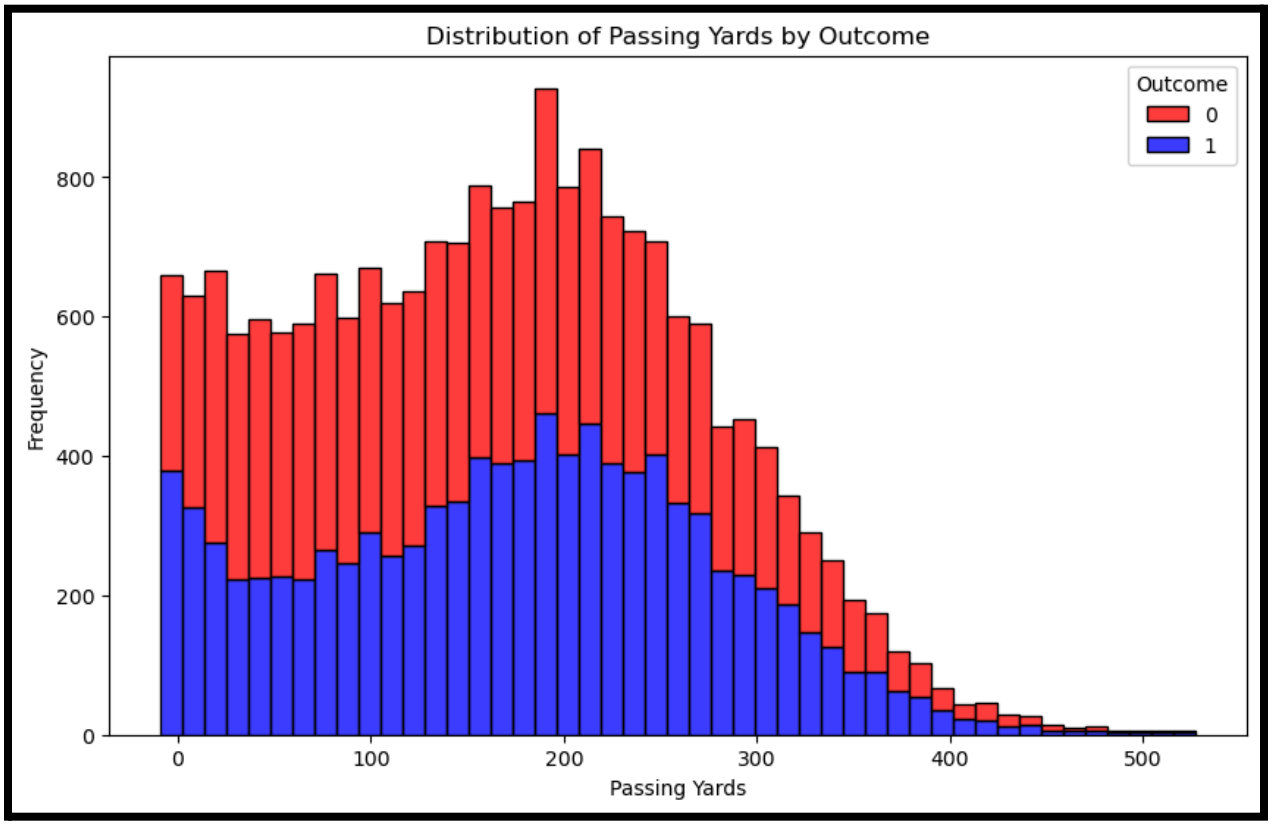


This analysis suggests that while yardage metrics are important, they may not be sufficient on their own to predict game outcomes, and other factors (possibly turnovers, defensive metrics, etc.) might need to be considered for a more accurate model.

Understanding the Distribution of the Data

This histogram shown in Figure 5a reflects the distribution of passing yards, with the data split by game outcome (win or loss). The blue bars represent games that were won (Outcome = 1), and the red bars represent games that were lost (Outcome = 0).

Figure 5a: Distribution of Passing Yards by Outcome



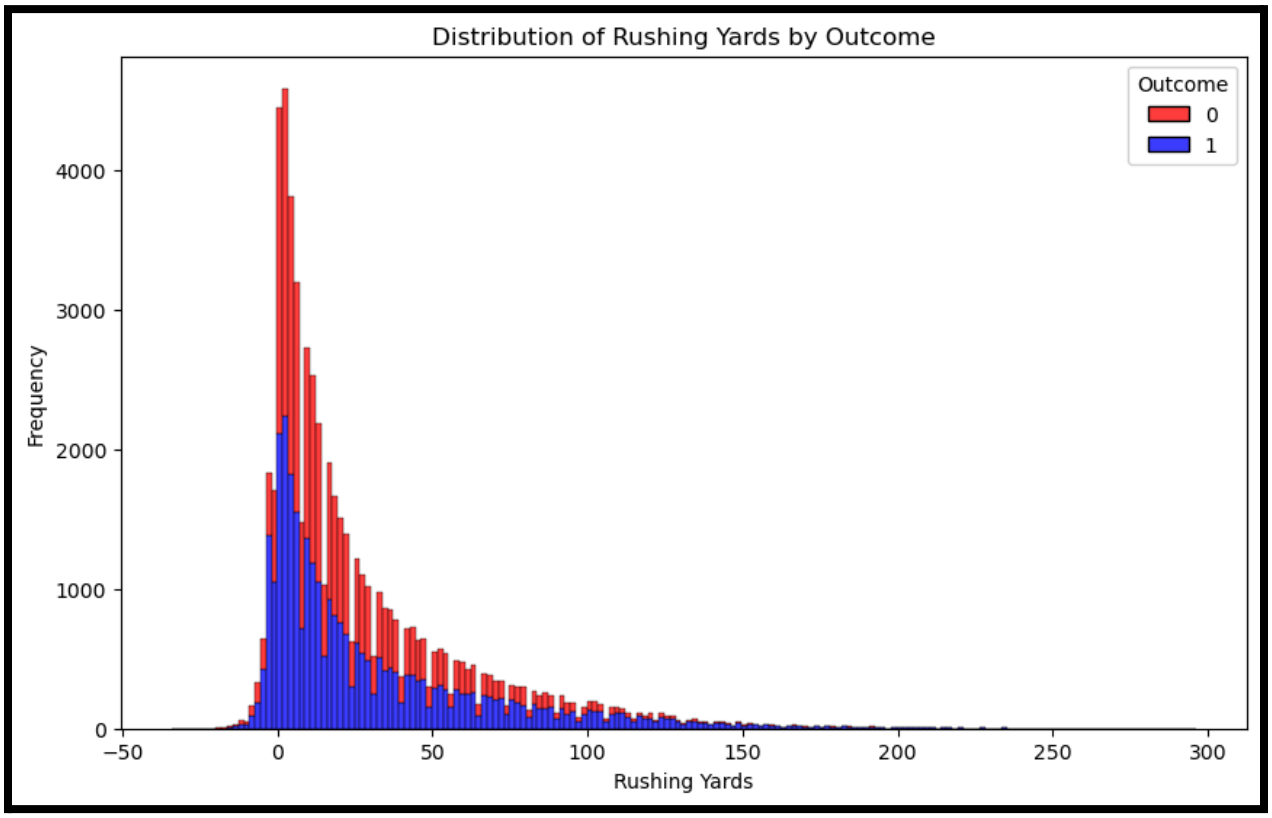
Key Observations:

- The distribution of passing yards is skewed to the right, with most games having fewer passing yards, and the frequency decreasing as passing yards increase.
- In the lower range of passing yards (0-200 yards), the red bars (losses) dominate the distribution. This suggests that games with lower passing yards are more likely to result in losses.
- As passing yards increase, particularly in the 200-300 yard range, the frequency of wins (blue bars) begins to increase relative to losses. This indicates that games with passing yards in this range are more evenly split between wins and losses.
- In the higher passing yard ranges (300+ yards), the blue bars (wins) become more prominent, suggesting that games with very high passing yards are more likely to be wins.

- There seems to be a *tipping point around 200-250 passing yards* where the likelihood of winning increases. This is indicated by the increasing proportion of blue bars relative to red bars in this range.
    - Low Passing Yards: Games with lower passing yards (below 100-150 yards) are more frequently losses, suggesting that insufficient passing performance may contribute to a team's loss.
    - Moderate to High Passing Yards: Games with passing yards in the range of 200-300 yards are more balanced between wins and losses, indicating that while higher passing yards may help, they are not a guaranteed indicator of a win.
    - Very High Passing Yards: Games with passing yards above 300 yards are more often wins, indicating that strong passing performance is likely correlated with winning.

Figure 5a suggests a clear relationship between passing yards and game outcomes. Lower passing yards are more strongly associated with losses, while higher passing yards, particularly above 300 yards, are more likely to result in wins. This insight could be valuable for understanding the importance of passing performance in determining game outcomes and could inform strategies for improving team performance.

Figure 5b: Distribution of Rushing Yards by Outcome



This histogram in Figure 5b shows the distribution of rushing yards, with the data split by game outcome (win or loss). The blue bars represent games that were won (Outcome = 1), and the red bars represent games that were lost (Outcome = 0).

Key Observations:

- The distribution is heavily skewed to the right, with most games having relatively low rushing yards. The frequency of games decreases rapidly as rushing yards increase.
- Comparison Between Wins and Losses:
    - Low Rushing Yards (0-50 yards): The majority of games with very low rushing yards (0-50 yards) are losses (red bars). This suggests that games with very poor rushing performance are more likely to result in a loss.
    - Moderate Rushing Yards (50-100 yards): As rushing yards increase into the 50-100 yard range, the proportion of wins (blue bars) starts to increase, though losses still dominate.
    - Higher Rushing Yards (100+ yards): In games with higher rushing yards (over 100 yards), the frequency of wins increases relative to losses. This suggests that higher rushing yard totals may be more

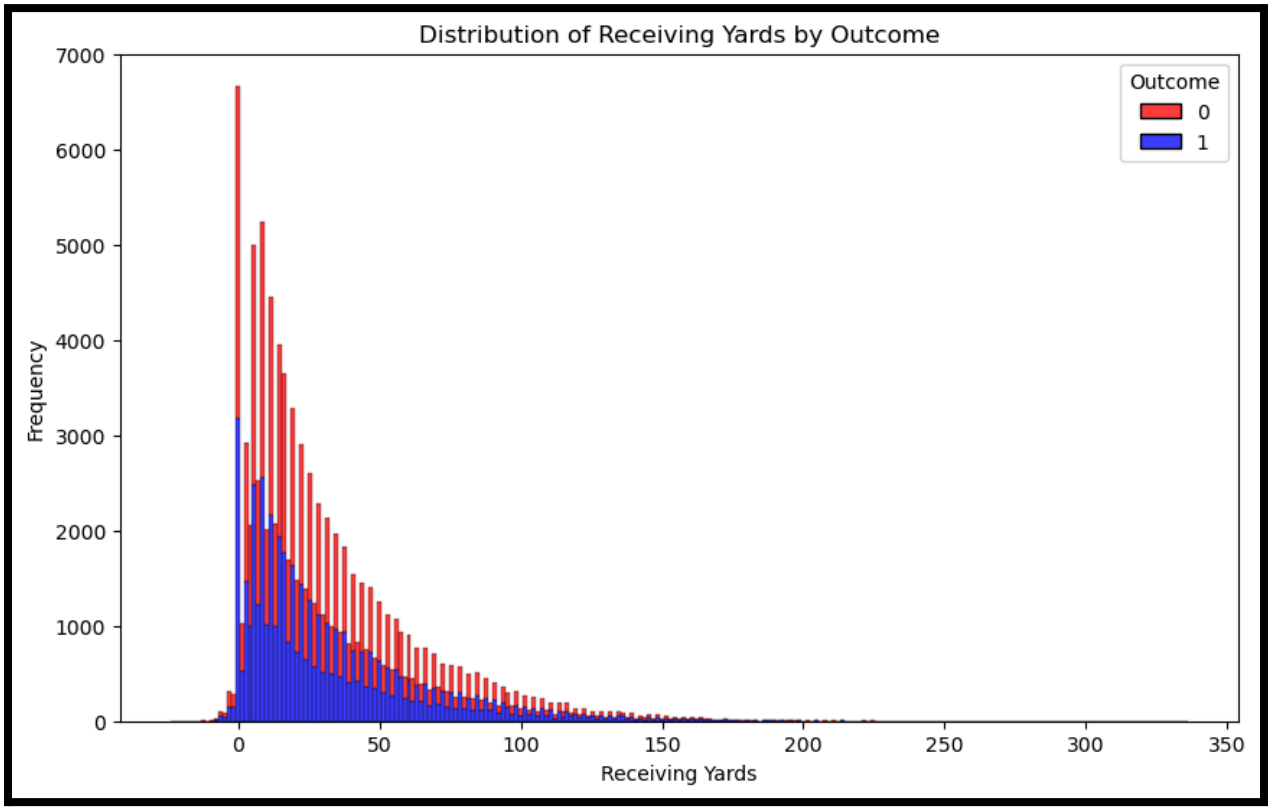
closely associated with winning outcomes, though these occurrences are less common overall.

- There is a noticeable *shift in the balance between wins and losses around 50-100 rushing yards*. As rushing yards increase beyond this point, the likelihood of winning appears to improve.
    - Very Low Rushing Yards: Games with very low rushing yards (close to 0) are overwhelmingly losses. This indicates that failing to generate rushing yards is strongly associated with losing.
    - Moderate Rushing Yards: The transition from losses to wins becomes more apparent as rushing yards move into the moderate range (50-100 yards). While not a guarantee of winning, moderate rushing yards seem to give teams a better chance of success.
    - High Rushing Yards: High rushing yards (over 100 yards) are less frequent but are more often associated with wins, suggesting that strong rushing performance is a key factor in winning games.

This Rushing Yards histogram highlights the importance of rushing yards in determining game outcomes. Games with very low rushing yards are much more likely to be losses, while higher rushing yard totals (especially above 100 yards) are more often seen in wins. The distribution suggests that while rushing yards alone may not guarantee victory, they play a significant role in a team's success, especially when they exceed certain thresholds.

This Figure 5b displays the distribution of receiving yards, with the data categorized by game outcome (win or loss). The blue bars represent games that were won (Outcome = 1), and the red bars represent games that were lost (Outcome = 0).

Figure 5c: Distribution of Receiving Yards by Outcome



Key Observations:

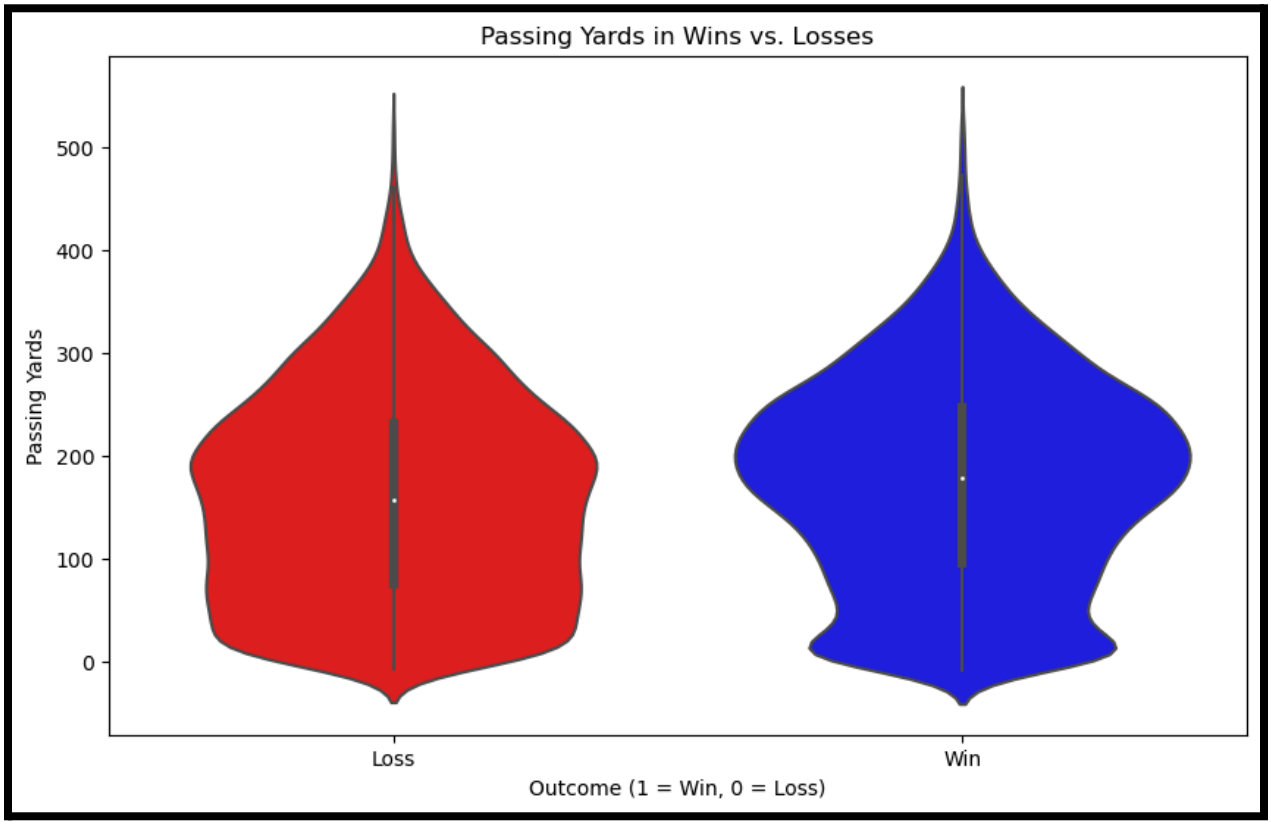
- The distribution is heavily right-skewed, with most games having relatively low receiving yards. The frequency decreases sharply as receiving yards increase.
- There is a high concentration of games with receiving yards close to 0, and the frequency drops off significantly as receiving yards increase beyond 50 yards.
- Comparison Between Wins and Losses:
    - Very Low Receiving Yards (0-20 yards): The majority of games with very low receiving yards are losses (red bars). This suggests that games with minimal receiving yards are more likely to result in losses.
    - Low to Moderate Receiving Yards (20-50 yards): As receiving yards increase slightly, the proportion of wins (blue bars) starts to increase, but losses still dominate this range.
    - Higher Receiving Yards (50+ yards): In games with higher receiving yards, the frequency of wins increases relative to losses. This suggests that games with higher receiving yard totals are more often associated with winning outcomes, although these occurrences are less frequent.

- There is a noticeable shift in the balance between wins and losses as receiving yards increase. *As receiving yards move beyond 50 yards, the likelihood of winning appears to improve*.
    - Very Low Receiving Yards: Games with very low receiving yards (close to 0) are overwhelmingly losses. This indicates that failing to generate significant receiving yards is strongly associated with losing.
    - Low to Moderate Receiving Yards: The transition from losses to wins becomes more apparent as receiving yards move into the moderate range (20-50 yards). While not a guarantee of winning, this range seems to give teams a better chance of success.
    - High Receiving Yards: High receiving yards (above 50 yards) are less frequent but are more often associated with wins, suggesting that strong receiving performance is a key factor in winning games.

Figure 5c highlights the importance of receiving yards in determining game outcomes. Games with very low receiving yards are much more likely to be losses, while higher receiving yard totals (especially above 50 yards) are more often seen in wins. The distribution suggests that *while receiving yards alone may not guarantee victory, they play a significant role in a team's success*, especially when they exceed certain thresholds.

How is the Data Distributed Between Wins and Losses?

Figure 6a: Passing Yards in Wins and Losses



The violin plot in Figure 6a provides a visual comparison of the distribution of passing yards between games that were wins (on the right, in blue) and losses (on the left, in red). This plot tells us:

Key Observations:

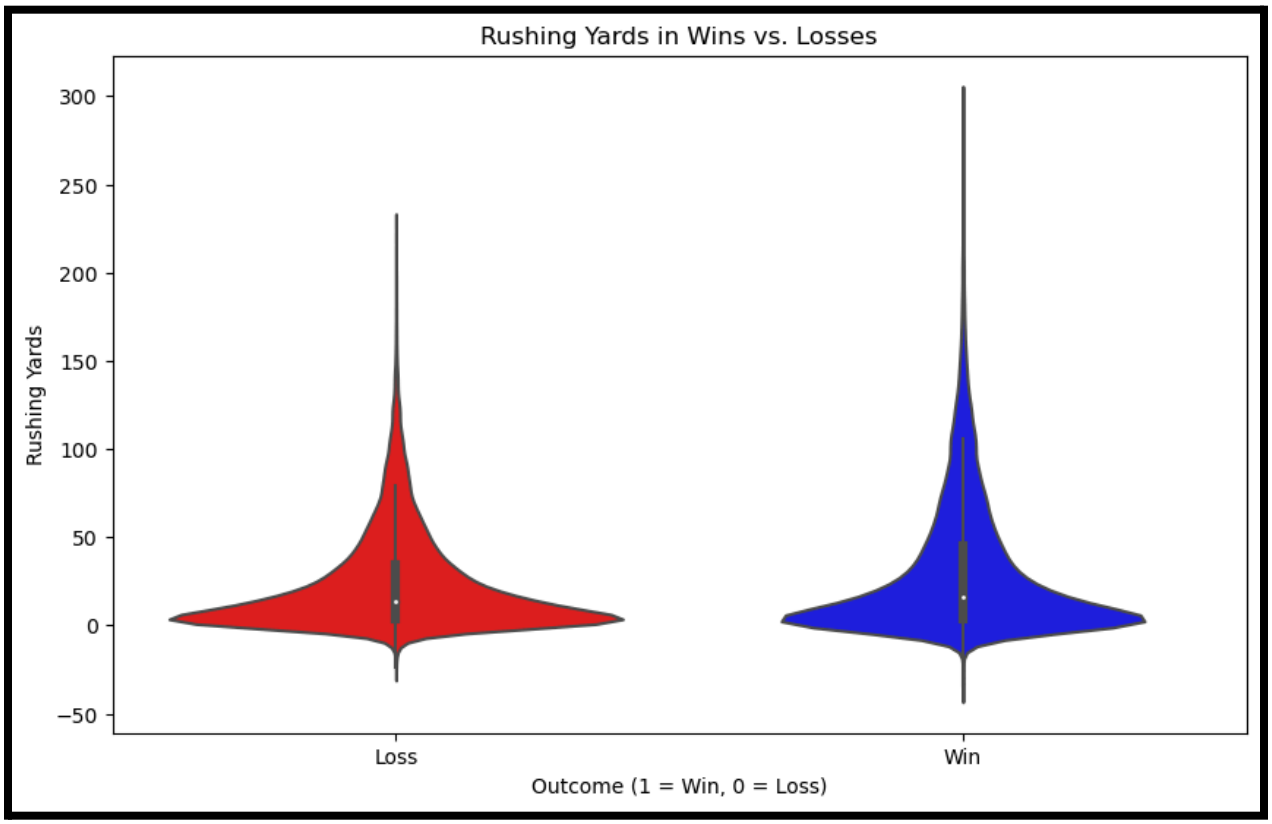
- Both wins and losses show a similar overall distribution shape, with the majority of games having passing yards clustered between approximately 100 and 300 yards.
- The median passing yards for both wins and losses appear to be similar, likely between 150 to 200 yards.
- Both distributions have some extreme values, with passing yards going above 400 or 500 yards in some games.
- The plot shows that while the overall shape and central tendency are similar, there is only a subtle difference in how passing yards spread out, particularly in the tails of the distribution.

- The density of the passing yards for losses (red) and wins (blue) seems fairly similar, suggesting that passing yards alone may not be a strong differentiator between winning and losing games.

This figure helps us understand the distribution of passing yards in wins versus losses, but it also suggests that passing yards alone might not fully explain the difference between winning and losing games.

There are subtle differences in the tails of the distribution, which could indicate that extreme passing yard values—either very high or very low—might influence the game's outcome differently. This aligns with the idea that in blowout games, a team may accumulate many yards if they are dominating, or conversely, have very few yards if they are being overwhelmed.

Figure 6b: Rushing Yards in Wins and Losses



This violin plot in 6b, above, compares the distribution of rushing yards between games that were losses (in red) and wins (in blue). This plot tells us:

Key Observations:

- Both distributions show that the majority of games have relatively low rushing yards, clustered between 0 and 50 yards.
- The median rushing yards appear to be slightly higher for wins than for losses, though the difference isn't substantial.
- The distribution of rushing yards for losses has a broader spread below the median, indicating a wider variability in low rushing yardage during losses.
- For wins, the distribution is slightly more concentrated around the median, with fewer extreme low values.
- There are also some higher rushing yards in wins, with values reaching above 150 yards, though these are less frequent.
- Both wins and losses have a significant concentration of games with low rushing yards, but wins show a slightly broader range of higher rushing yard values.
- This might suggest that while rushing yards aren't drastically different between wins and losses, games with higher rushing yards are somewhat more associated with wins.
- The plot suggests that having higher rushing yards could be slightly more indicative of a win, but the overall impact appears to be moderate.

This violin plot provides insights into how rushing yards are distributed across games with different outcomes, indicating that while there is some correlation between rushing yards and winning, it's not a strong or definitive factor.

Figure 6c: Receiving Yards in Wins and Losses

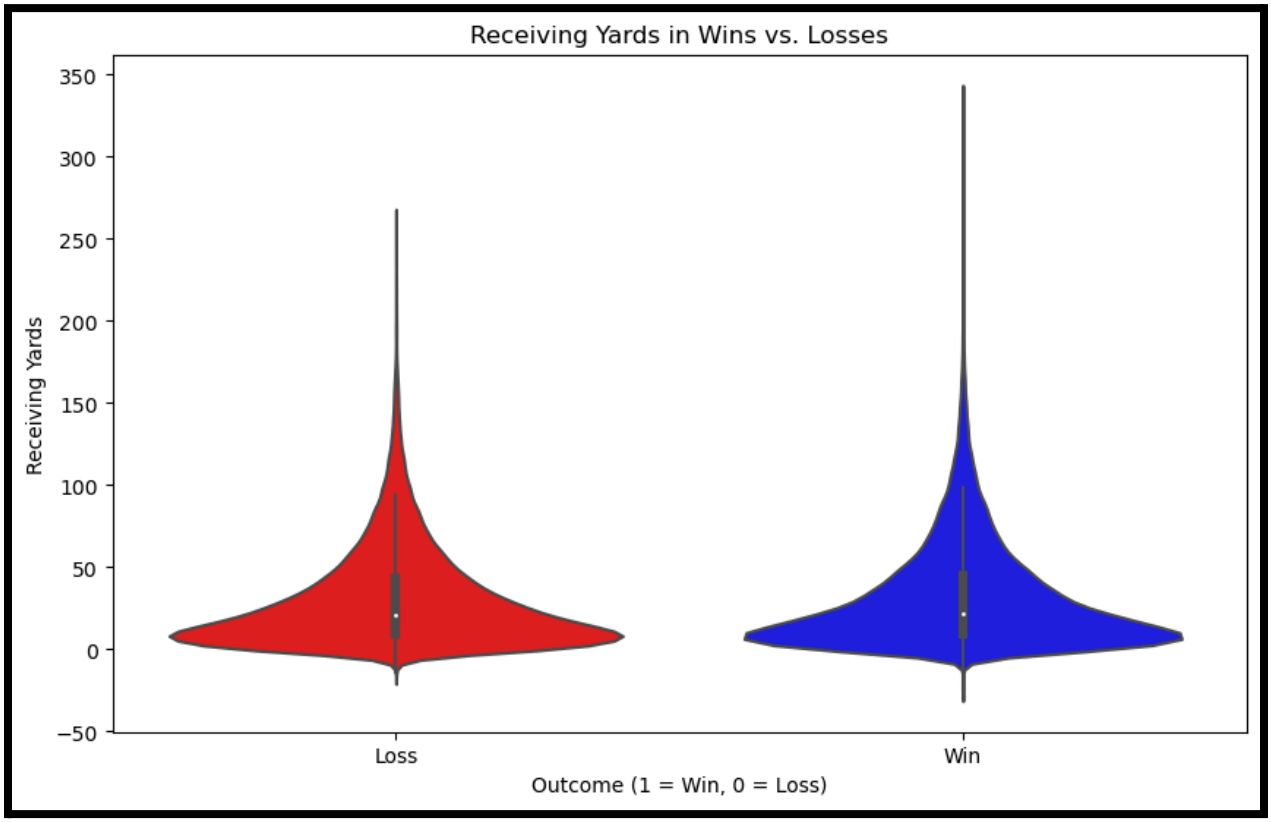


Figure 6c compares the distribution of receiving yards between games that resulted in losses (in red) and wins (in blue).

Key Observations:

- Most games have relatively low receiving yards, with a large concentration between 0 and 50 yards.
- The median receiving yards appear to be quite similar for both wins and losses, indicating that receiving yards might not be a strong differentiator between winning and losing games.
- The distribution for wins shows a slightly broader range of higher receiving yard values compared to losses, which suggests that higher receiving yards could be slightly more associated with wins, though the effect doesn't appear to be strong.
- There is a slight indication that games with higher receiving yards might be more common in wins than in losses, but the difference is not substantial.

- The distribution of receiving yards is generally similar for both wins and losses, with most games having low receiving yards. However, wins show a slightly broader range of higher receiving yard values.
- While receiving yards might have some impact on the game's outcome, the plot suggests that they are not a decisive factor, as the distributions for wins and losses are quite similar.

The violin plot in Figure 6c provides insights into how receiving yards are distributed across games with different outcomes, indicating that while there may be a slight association between higher receiving yards and wins, it is not a strong or clear-cut relationship.

Explaining the Differences

One glaring difference in the charts is between the Random Forest and the Histogram of Passing, Rushing, and Receiving Yards. There can be many reasons why these two charts seem to be in conflict. One key reason is that the Random Forest model captures complex, non-linear relationships and interactions between features that a simple histogram does not. While the histogram may suggest a straightforward linear relationship between these yardage metrics and game outcomes, the Random Forest model may determine that these metrics are less important when considered in the broader context of all the features. The Random Forest considers how variables interact with each other and may identify other factors as more critical to predicting outcomes, thereby reducing the perceived importance of the passing, rushing, and receiving yards.

Another possible explanation for the difference is the presence of multicollinearity or redundant information among the features. The histogram examines each feature in isolation, potentially highlighting a correlation that appears strong when not accounting for the influence of other variables. However, the Random Forest, which evaluates the combined effects of multiple variables, may find that the importance of one yardage metric diminishes when similar or related metrics are also included in the model. Additionally, the Random Forest model may highlight that certain ranges of yardage are more predictive of outcomes than others, while the histogram provides a broader, more general view that could overlook these nuances. These differences underscore the importance of using multiple analytical approaches to gain a comprehensive understanding of the data.