Exploring the Correlation between Skill Position Yards and Home Advantage on Football Game Outcomes

Names: Nathan Kriz, Harshet Anand, George Richardson

Introduction

Football is a sport that attracts millions of fans worldwide, us being interested in the NFL. With such a following comes a natural curiosity to understand the factors that contribute to a team's success. One such factor that has been frequently analyzed is the number of yards that a team's skill positions acquire during a game. In this data analysis project, we seek to explore whether there is a correlation between a team gaining a home advantage by analyzing many of the games' scores and looking at the number of games won by a team playing at home and a team playing away. We will also be looking at a team's total yards gained by their skill positions and their likelihood of winning the game by comparing the stats of each position and the number of points scored by each team. By answering this question, we hope to provide insights into what factors are essential in securing a win in football and contribute to the existing body of research on the sport. Our analysis throughout this project will comprise a regression analysis to determine the impact of yards and points on winning a football game, as well as a binary logistic regression to deal with binary outcomes and the effect of different variables present when looking at the impact of playing at home. Lastly, we will be focusing on regression lines that explore the relationship between yards and wins as well as points scored and wins. This will be central to the main question, which focuses on the following:

- What is the ability to predict the outcome of NFL games based on the statistics of how the team performed in a game when playing at home and the usage of certain skill positions, and the impact of each skill position on a team?

Background

Information about the Dataset

The dataset which we will use for our project was put together by the website advanced sports analytics. This data was collected by compiling the stats of all offensive skill positions for every team in the NFL for all games that were played. We selected data from the 2019-2023 NFL seasons.

Some key variables which are present in this dataset include pass completions, passing yards, passing touchdowns, interceptions, rush yards, receiving yards, rushing touchdowns, and receiving touchdowns. When focusing on home advantage, some key variables present are home score, away score, and which teams played at home. All of these variables which are in our dataset are statistics that are tracked during each NFL game as they can show whether a player performed well or not in the football game, especially when comparing if they played home or away. Since this dataset comes from NFL statistics from 2019-2023, there is no larger population that this data comes from since this dataset has every game in each season. A lot of these variables will come to use when analyzing the impact of home advantage as we will be able to see whether home advantage makes an impact on every team or only a certain number of teams.

The data set we are using for this analysis presents multiple variables which will help us be able to determine the impact between skills positions and home advantage on the impact of winning games. If there is a high percentage of games won, this leads to a higher probability of winning a football game when playing at home. Additionally, if a certain team utilizes their skills positions better when comparing the number of yards each team has gained with each position, this will also lead to a higher probability of winning a football game.

This data set was collected by analyzing every single football game from the past four years between 2019 and 2023. Each variable is important in helping us determine the impact and importance of factors that contribute to a higher percentage of a team winning. The most important variables that will play a role include every variable that focuses on the number of yards gained by each player and their respective positions, as well as the home team and away team scores and whether or not they were playing at home or away.

Data Source

Citation: Advanced sports analytics. (n.d.). Retrieved April 24, 2023, from https://www.advancedsportsanalytics.com/nfl-raw-data

Background Information

There are some factors that would play a big role in determining whether or not home advantage has a significant impact on the probability of a team winning a game or not. Some background information we will need to consider to better comprehend our analysis is which team is playing home and which team is playing away. Some other factors that we would have to consider are receiving yards and rushing yards, and whether or not if more yards are gained, this would increase the chances of winning a game. We will have to look at if certain positions gain more yards and the impact on whether or not this factor would increase or decrease the percentage of winning a game by potentially looking at some of the highest-scoring games over the past four years.

Effect of Unusual Factors

Some unusual factors which could affect our analysis and our interpretation of our results are the fact that some stats are insignificant in the game. For example, when one team is winning by a large margin, the other team may put up large stats late in the game, which would have no effect on the outcome, so this could impact our analysis, though it shouldn't be a major issue for us. There are also some games that end in a tie, which does not count as a win for either team, so these data points are not useful in our analysis. Although many of these games go into overtime, the data present whether a team won in overtime or not. Thus, although a game ending in a tie may present as an unusual factor for certain games, we can analyze the impact of overtime and see the impact that this factor has on home advantage, so this would have to be an unusual factor to consider and analyze. Another unusual factor we will have to consider when looking at our data is weather conditions. Weather can have a big impact on a football game's result and may even interact with the connection between skill position yards and home-field advantage. For instance, if a team is playing in bad weather, it might be harder for its skill-position players to get yards, which might make the link between yards and home advantage less intense. Lastly, since our dataset only has data for seasons from 2019 to 2022, with the end of the 2022 season going into 2023, there is much less game data for the year 2023, so it is not as reliable as the other years which have a full year of game data.

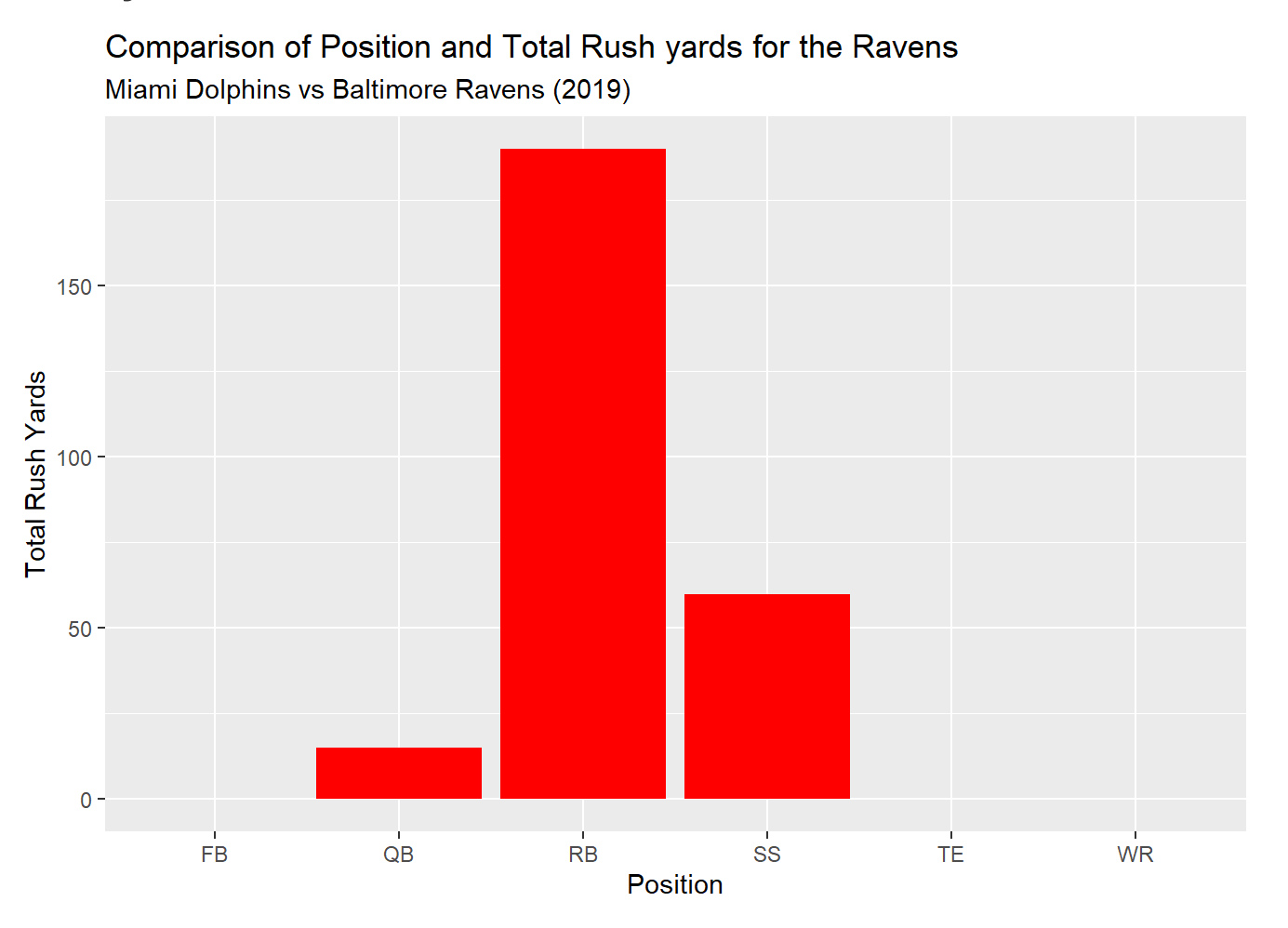
When analyzing skill position for each team and the number of yards they gain throughout the game, there are no unusual factors that we would have to consider, as any factors that are different between certain teams would be the factors that would prove that skill positions and the yards gained do in fact have an impact on the outcome of a game.

Focus of the Report

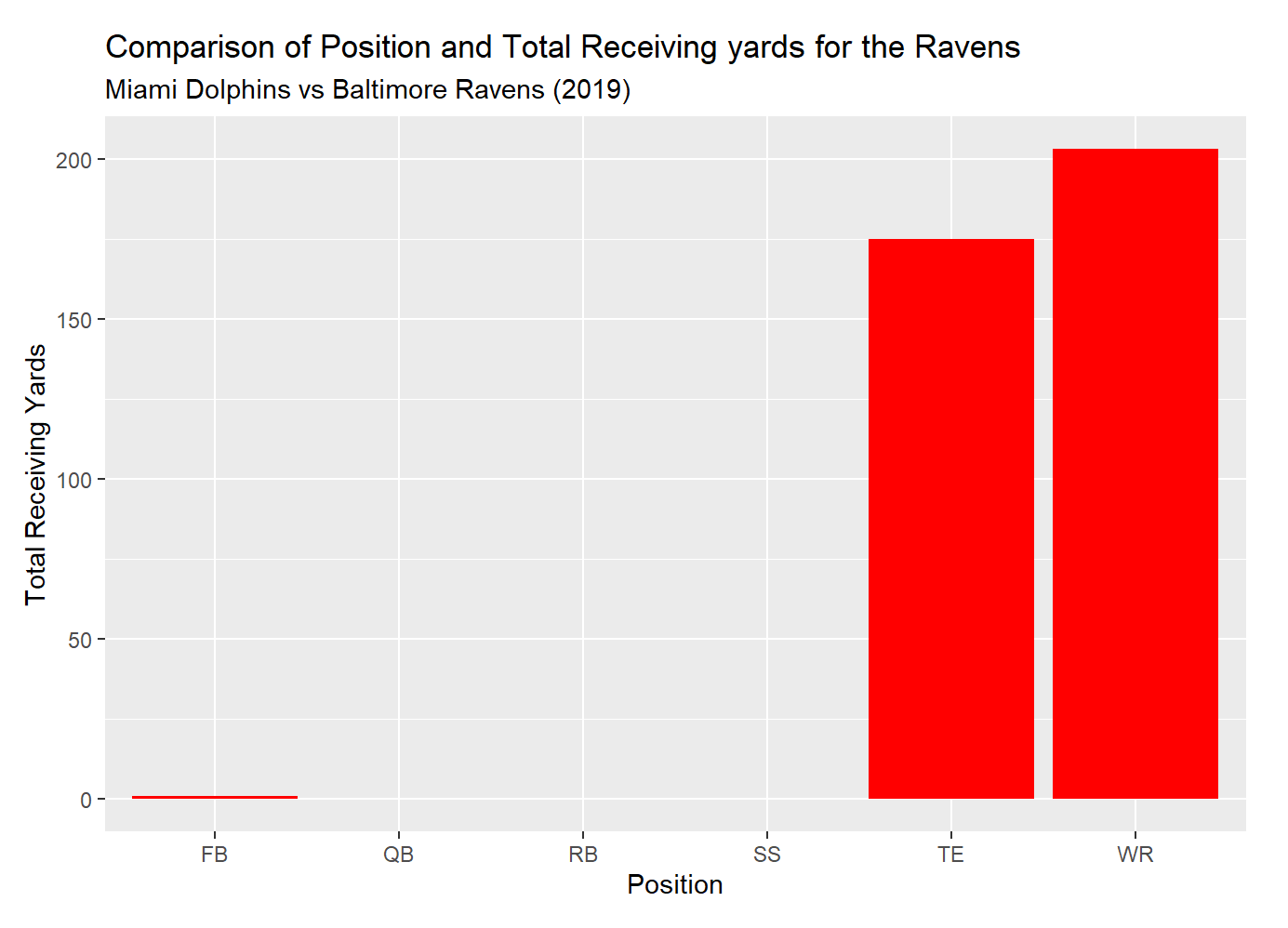
For the rest of the report, we will create visuals and graphs which show different stats for different skill positions and how they relate to the outcome for their team, and whether they won the football game or not. We will also add regression lines that explore the relationship between yards and wins in addition to points scored and wins.

The focus of the report is mainly on whether a team playing at home would have any sort of impact or a significant impact on determining the outcome of a game based on the statistics and data provided to us rather than basing it solely on the fact that they are playing at home for example solely being based on more home fans and the psychological impact this can have on the players. If we are able to compare the data provided to us from every game within the past four years, our analysis will give a better understanding of providing us with the statistical information needed to be able to predict the outcome of a game. Additionally, we will be able to determine whether or not certain skill positions gaining yards will have an impact on a game outcome and if this increases or decreases the probability of a team winning.

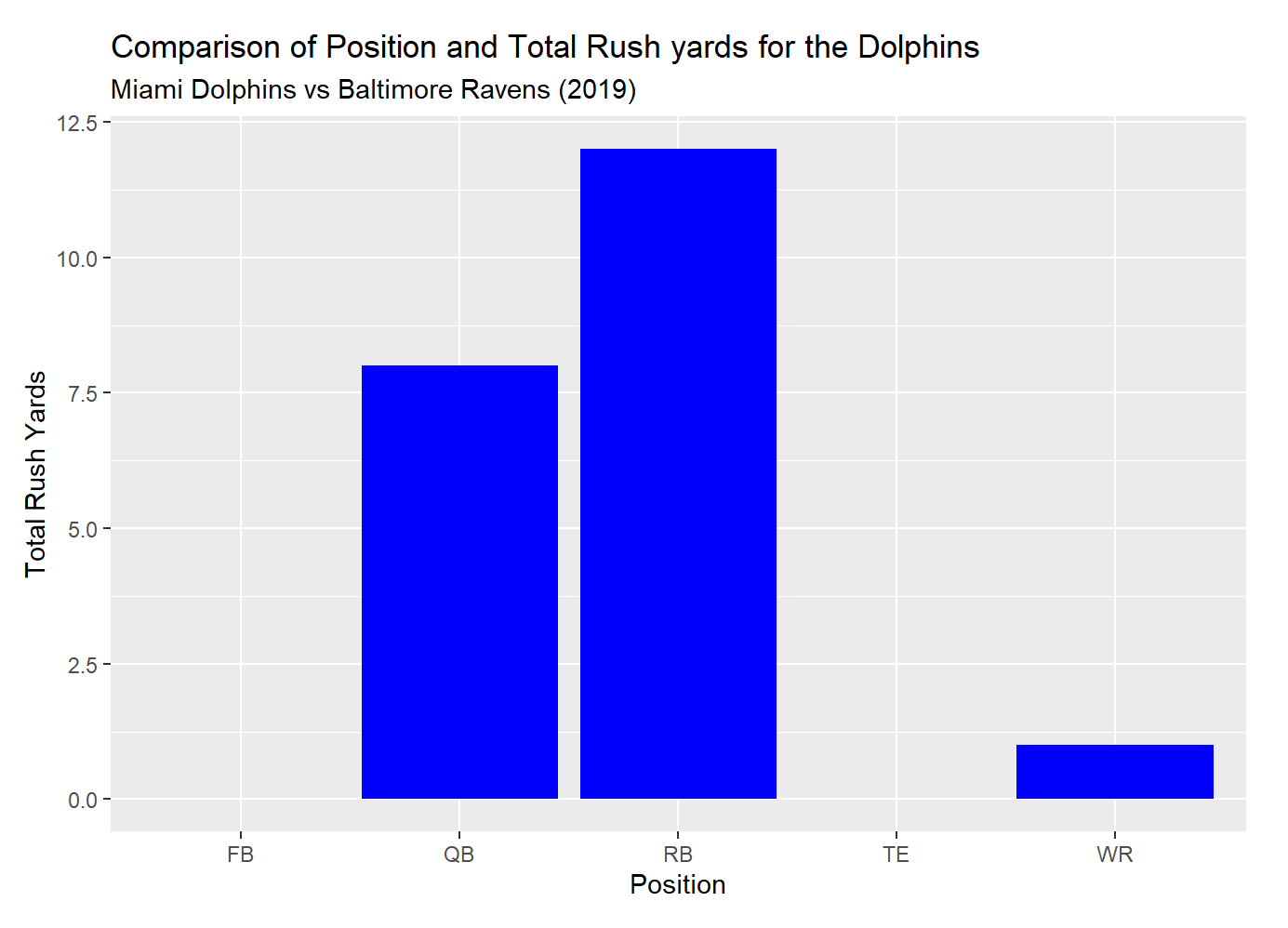
Analysis



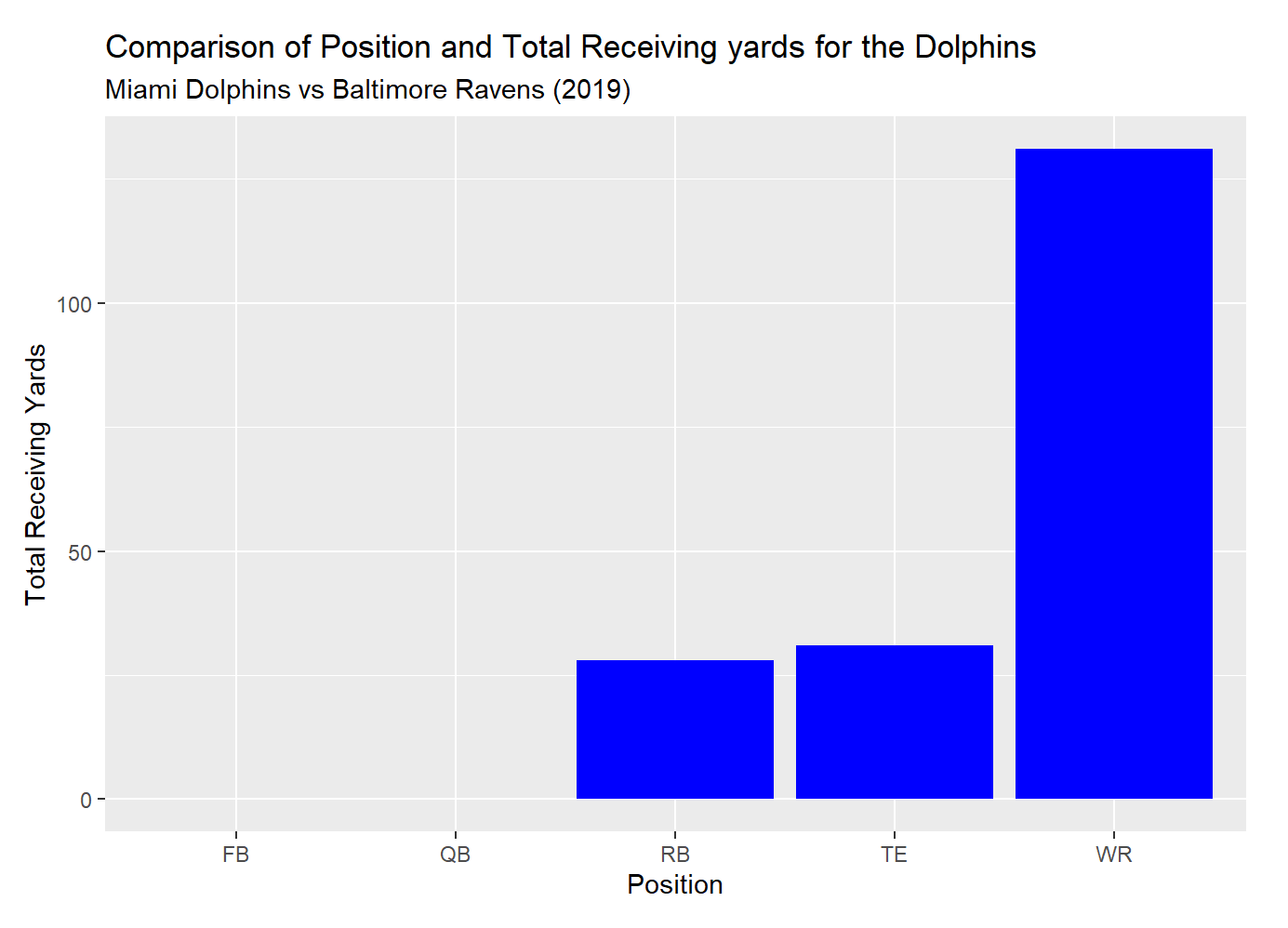
This histogram shows the total rush yards for team **Baltimore Ravens** in their football game against the Miami Dolphins and shows the rush yards for the RB, WR and SS positions. The SS position is not present for the Miami Dolphins which shows that this skill position was important to the outcome of the game.



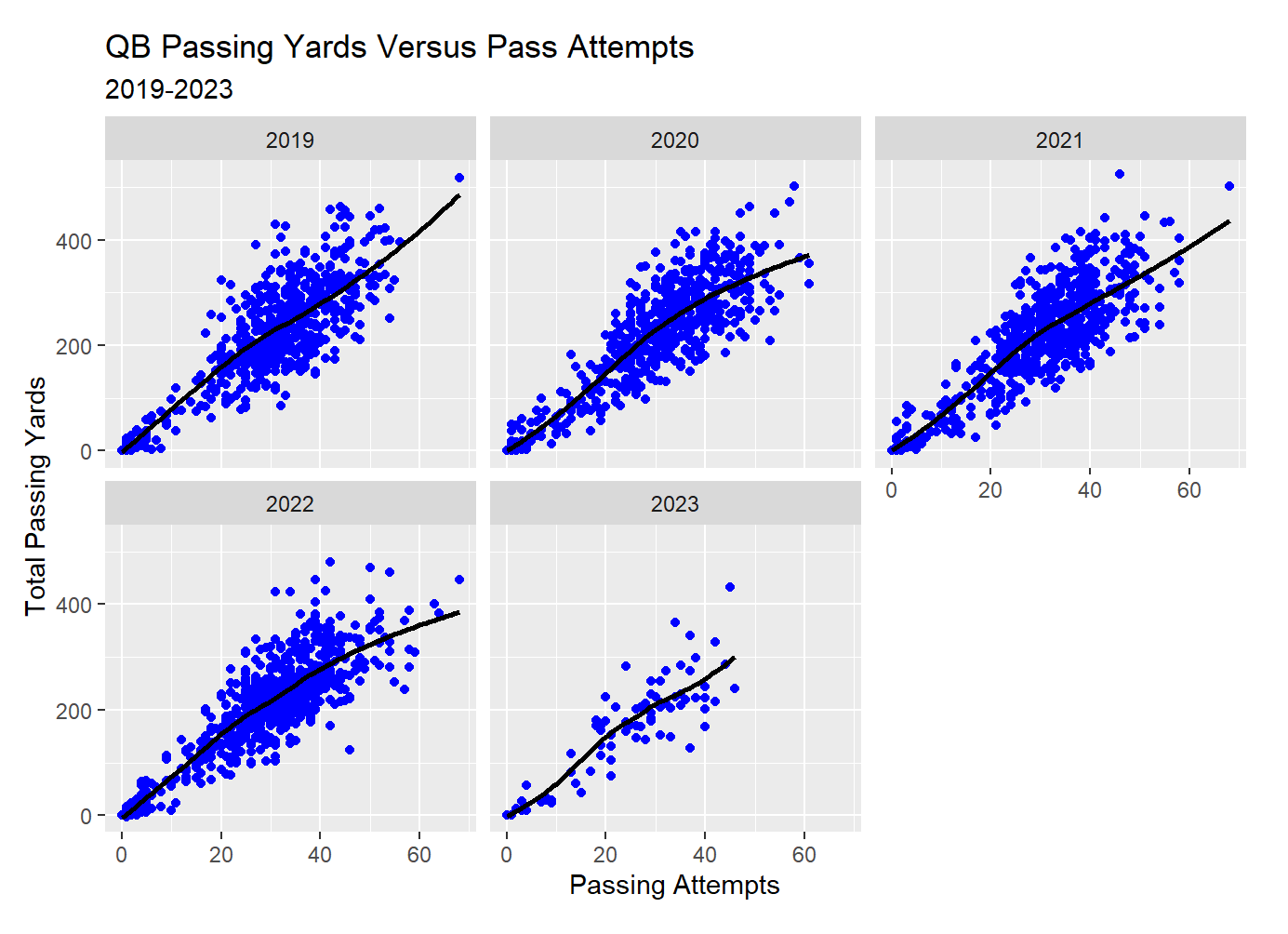
This histogram shows the total receiving yards for team **Baltimore Ravens** in their football game against the Miami Dolphins and shows the receiving yards for the WR, TE and FB positions. Both the WR and TE positions are most important skill positions for the Ravens.



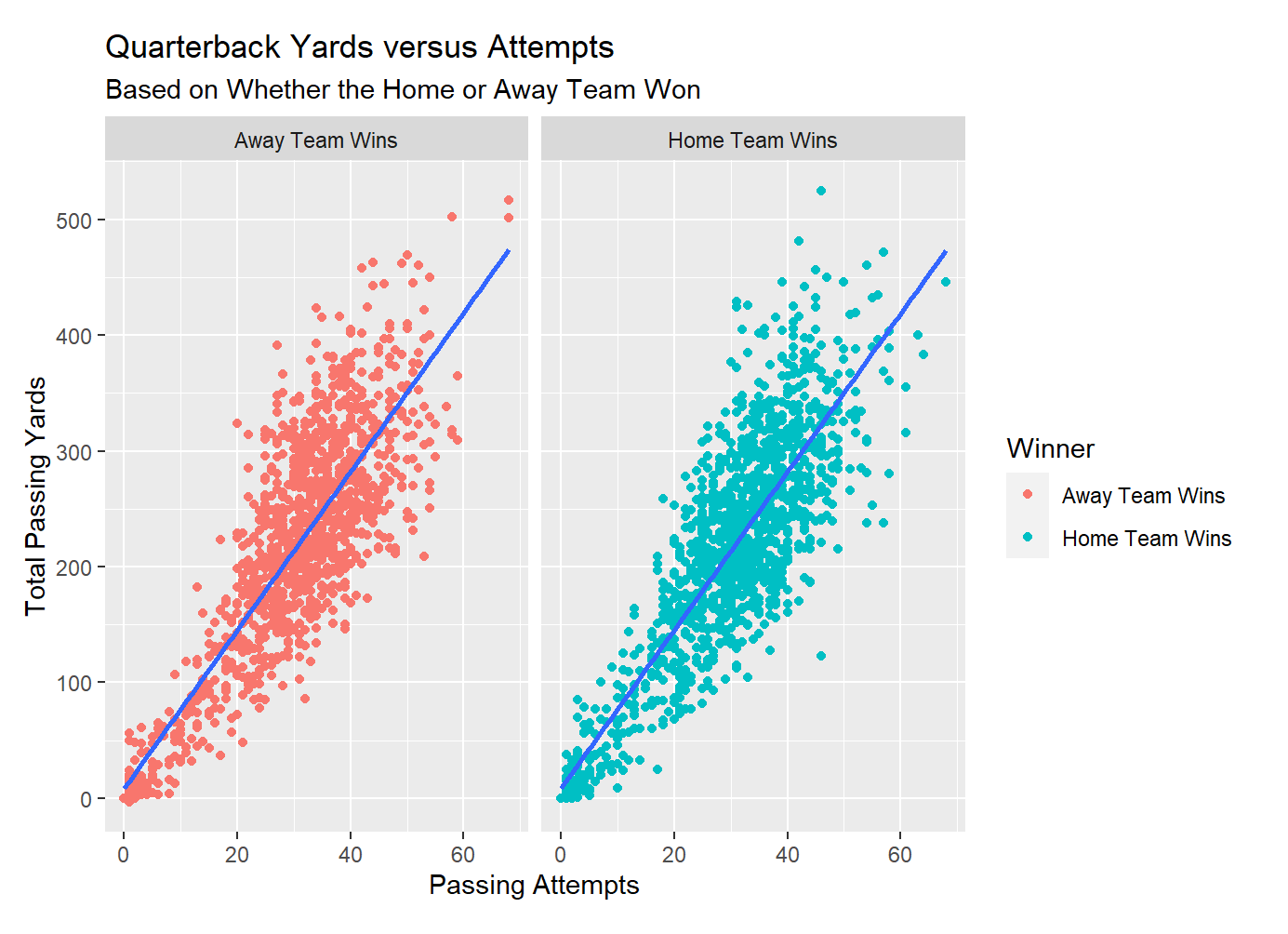
This histogram shows the total rush yards for team **Miami Dolphins** in their football game against the Miami Dolphins and shows the rush yards for the RB, WR and QB positions. For the Dolphins, the RB position is the most valuable position in rushing.



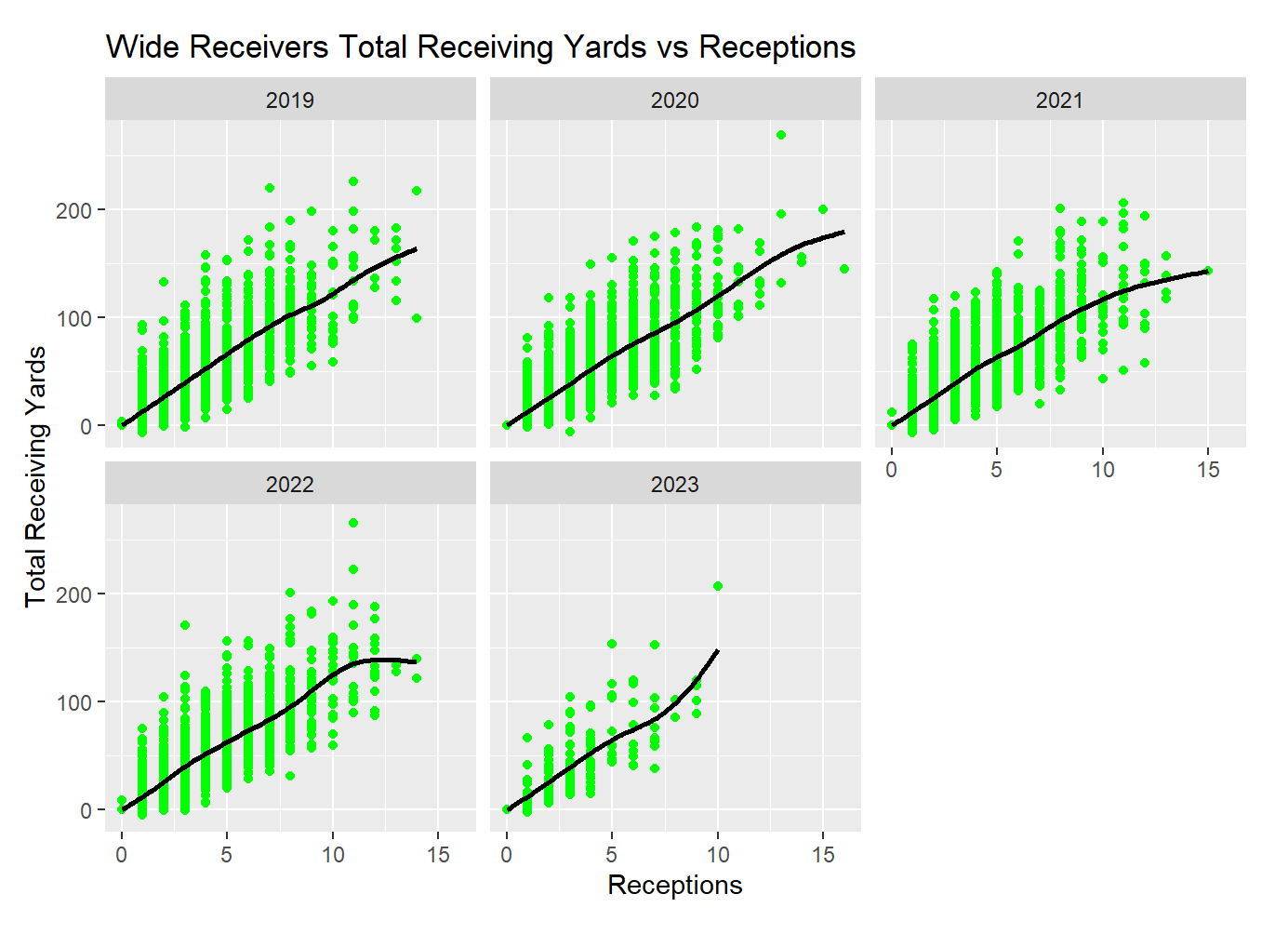
This histogram shows the total receiving yards for team **Miami Dolphins** in their football game against the Miami Dolphins and shows the rush yards for the WR, RB and TE positions. In receiving, the WR position is the most impactful for the Dolphins offense.



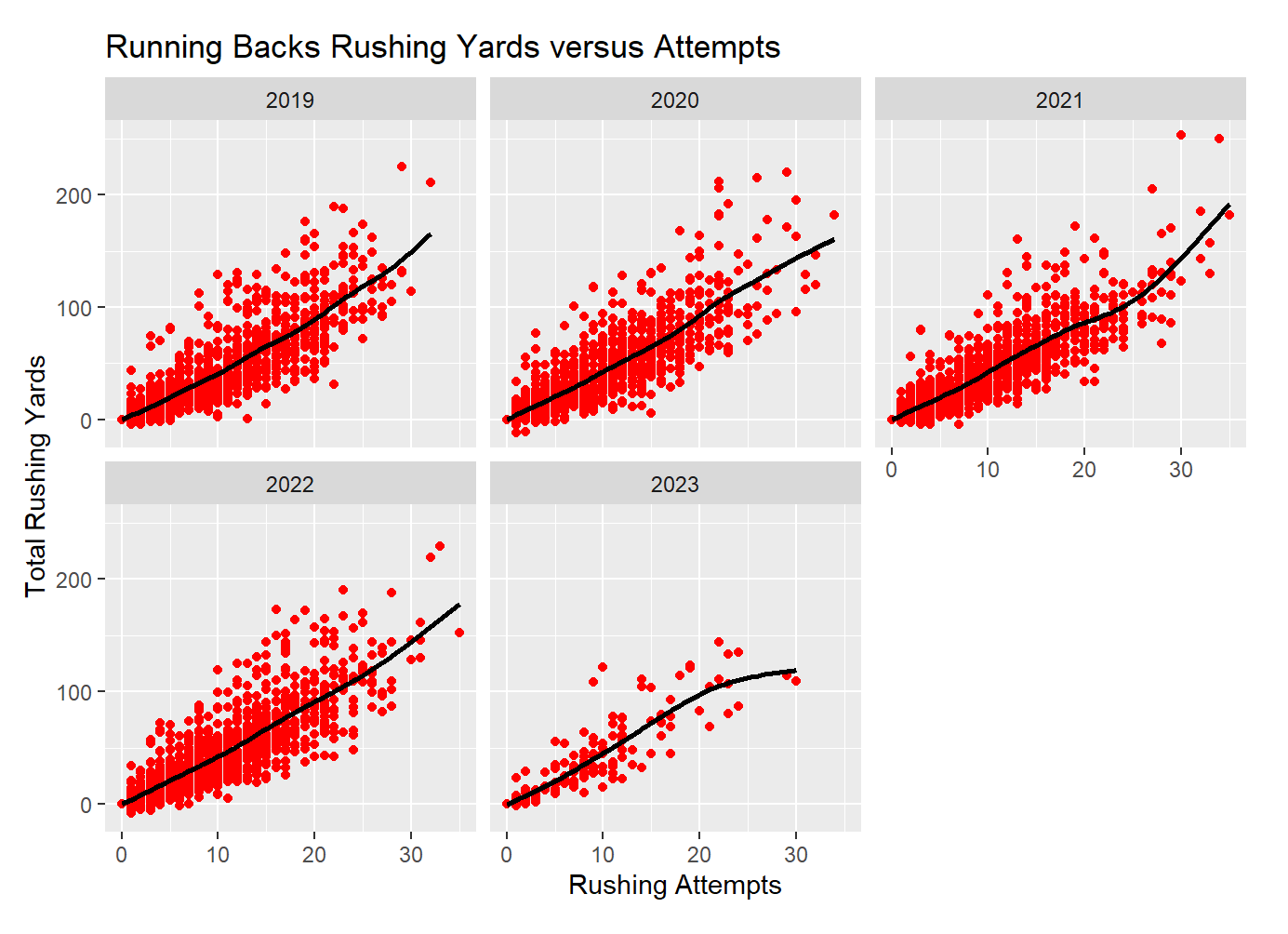
This set of scatter plots show the trends of QB passing yards over the years from 2019 to 2023 and shows the trend between the passing attempts of a QB and the total passing yards. The graphs show a trend which has stayed relatively the same throughout the time period, with some QBs throwing for more yards in the more recent years. The QB is the most influential position in determining the outcome of the game, so the more yards the QB throws, the better the chances of their team winning.



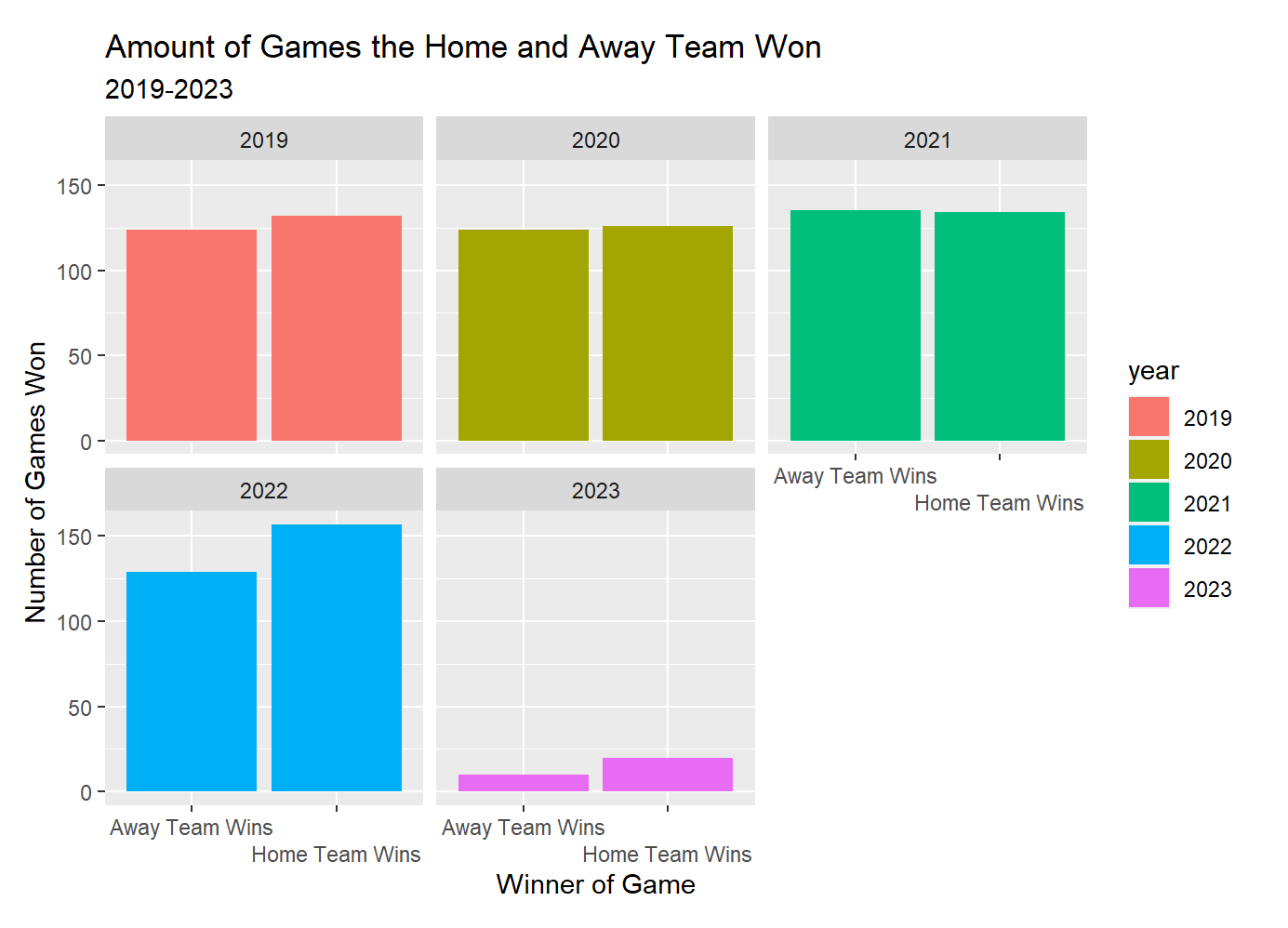
This graph shows a breakdown of the quarterback passing yards versus pass attempts based on whether the home team won the football game or the away team won the game. This graph demonstrates that there are more occurrences of the home team quarterback throwing for 300 plus yards which then resulted in their team winning the football game for both home and away teams.



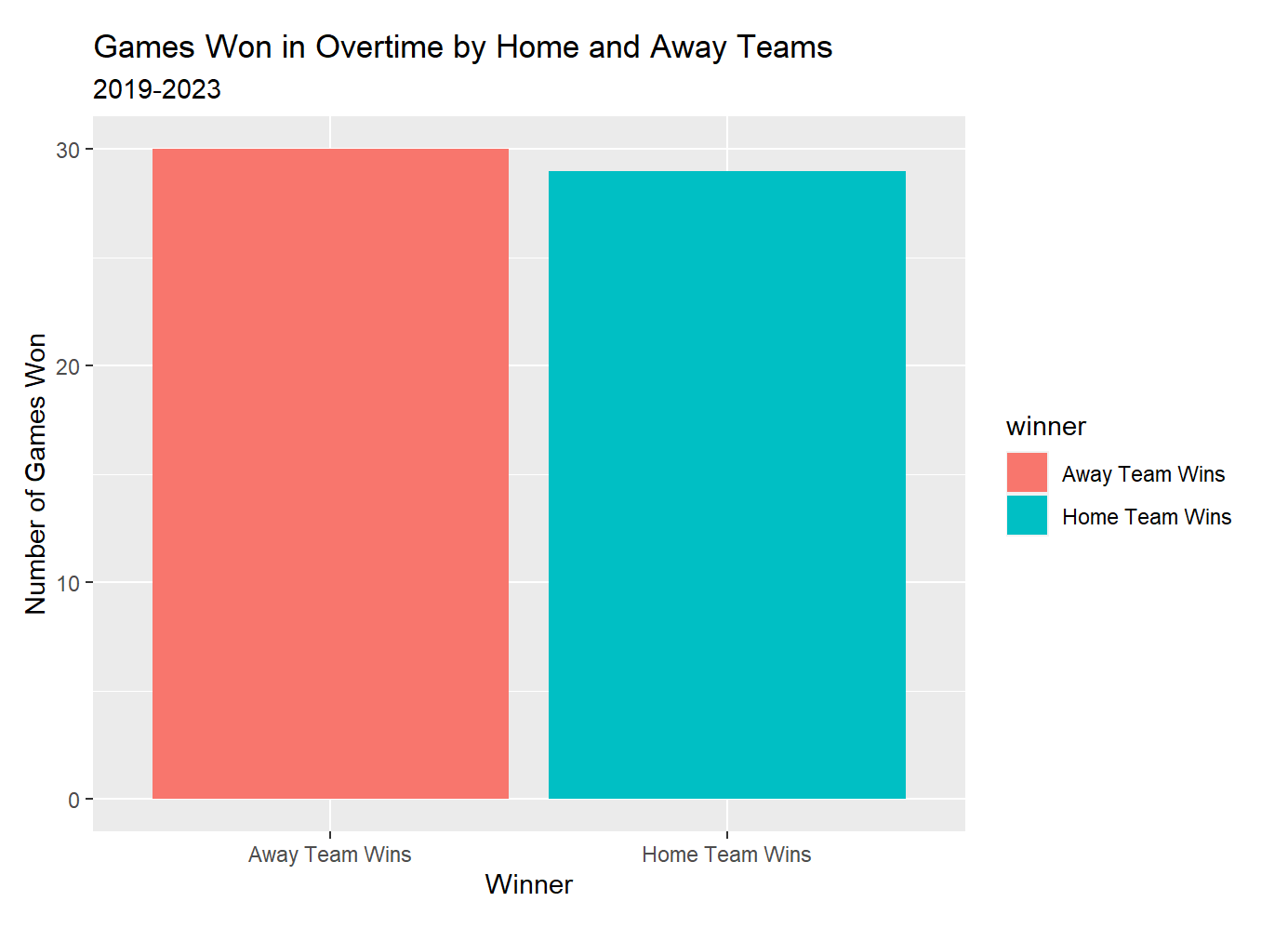
This graph shows all wide receivers for the years 2019-2023 and their receptions and receiving yards in each game. It can be seen that wide receivers account for many more yards than other skill positions other than the quarterback as the trend for wide receiver yards has slowly increased throughout the years examined.



This graph shows all running backs for the years 2019-2023 and their rushing attempts and rushing yards in each game. The data shows a small decrease in overall rushing yards by running backs from 2019-2023 as many teams are starting to rely less on running backs for yards, with other skill positions accounting for more of the yards.



This graph shows a breakdown by years from 2019-2023 which show whether the home team or the away team won each NFL game. This data shows that in general, the home team has a higher chance to win the football game, with years 2020 and 2021 being more even in terms of home teams and away teams winning the game.



This graph shows the games that went to overtime from 2019-2023, not including ties, and which team won the game, either the home team or the away team. This histogram demonstrates that there is not a clear difference in whether being the home team will lead to winning the game more often in overtime. Instead, it is relatively even once the game goes into overtime whether the home or away team will win the football game.

```
##
##  One Sample t-test
##
## data:  team_pts$home_score
## t = 78.851, df = 1086, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  23.06099 24.23800
## sample estimates:
## mean of x
##  23.64949
```

After finding a confidence interval for the home team score in NFL games from 2019-2023, we are 95% confident that the mean home team score is between 23.06 and 24.24 points.

```
##
##  One Sample t-test
##
## data:  team_pts$vis_score
## t = 74.42, df = 1086, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  22.04963 23.24384
## sample estimates:
## mean of x
##  22.64673
```

After finding a confidence interval for the away team score in NFL games from 2019-2023, we are 95% confident that the mean home team score is between 22.05 and 23.24 points.

It can be seen from our confidence intervals of home team scores and away team scores that there is a difference in average scores between the home team and the away team. We will now complete a hypothesis test using the home scores and away scores to see if this difference is statistically significant or not.

Hypothesis Test

$$D_i \sim F(\Delta, \sigma), \quad \text{for } i = 1, \ldots, n$$

- F is a generic distribution for the population of differences
- Δ is the mean of this distribution
- σ is the standard deviation

$H_0 : \Delta = 0$
$H_a : \Delta \neq 0$

For this hypothesis test, we decided to see if the difference between home scores and away scores was 0. So we set our null hypothesis to be $\Delta = 0$. We then decided to put our alternative hypothesis as $\Delta \neq 0$ as this would help us indicate if there is a clear difference between the home and away team scores as the difference would be not equal to 0.

- Test Statistic

$$T = \frac{\bar{d} - 0}{s/\sqrt{n}}$$

```
##
##  Welch Two Sample t-test
##
## data:  team_pts$home_score and team_pts$vis_score
## t = 2.3469, df = 2171.5, p-value = 0.01902
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1648563 1.8406635
## sample estimates:
## mean of x mean of y
##  23.64949  22.64673
```

For this hypothesis test, the two sets of values that were used are the list of all the scores for the home teams in NFL games from 2019-2023 and the other was also a list which contained all the scores of the visiting teams. After running the hypothesis test, we obtained a p value of 0.019 which is less than our alpha value of 0.05 which clearly demonstrates that the difference in means is not 0. This shows that there is a clear advantage for teams that are playing at home when looking at what team will win the football game.

Discussion

After conducting our analysis looking at NFL games and whether there is a correlation between when a team is home or away and the outcome of the game, it can be seen that there is a correlation between a team being the home team, and them winning the football game. This was initially confirmed in our analysis when looking at the histograms which counted the number of games the home teams won and the number of games the away team won. In the years 2019, 2020, 2022 and 2023 the home team won more games than the away team, and in 2021 the two totals were very similar with the away team winning a few more games. Next, the confidence intervals of the mean scores for both home and away teams revealed a similar conclusion as the home teams had scored on average 1 more point than the away teams throughout 2019-2023. Finally, after conducting the hypothesis test, it was found that the difference in means between home and away scores was not zero, and instead was statistically significant as the p value that was found was less than the alpha value, leading to the conclusion that home teams have a better chance at winning the football game compared to away teams between 2019-2023. Additionally, when looking at the statistics and comparing each variable and the impact of yards gained by each skill position has on the outcome of each game, there is a significant difference that goes to show that, in fact, skill position does have an impact on the performance of each team, thus affecting the outcome of the game. When looking back at the thesis, we can tell that certain teams have a certain position that is utilized much more efficiently compared to other teams, and this is much more apparent when looking at the number of yards each position has acquired.

One major shortcoming that was encountered was that there were a few games that ended in a tie even after overtime, and this had an impact as some of those games had certain statistics and important variables that would have been able to strengthen our analysis further and this would definitely be one aspect that would need to be further analyzed in the future. While the data we collected was a vast amount of games that were accumulated over the course of four years, some other methods we could use to be able to dig deeper into our analysis and the impact of our correlations are by comparing the possession each team has during the game in order to see if this has an impact on yardage thus further helping us determine the game outcome. The new data we would collect would have to correlate with the possession each team has gained in every game to be able to help determine the outcome of a game. One main question that we would further investigate would be, "Do certain types of yardage (e.g., rushing yards vs. receiving yards) have a stronger relationship with game outcomes, and does this relationship vary by position?" Although there are different types of yardage that can be earned throughout a football game, we can tell that when certain skill positions earn more yardage than others, there is a huge impact that it can have on determining the outcome of a game by noticing that some of the biggest differences between scores in football games are achieved by various skill positions being utilized effectively and those certain positions earning more yardage for each team.