

2020-0108 IST 652 Scripting for Data Analysis

Amelia Green, Andrea Smith, Will Holt

Submitted:

March 19, 2020

Table of Contents

# Contents

[Background and Questions 3](#_Toc35539739)

[The Data 3](#_Toc35539740)

[NFL\_SeasonResults 4](#_Toc35539741)

[SeasonSummary\_1970\_2019 4](#_Toc35539742)

[Modern 4](#_Toc35539743)

[The analyses 4](#_Toc35539744)

[Final Thoughts and Next Steps 12](#_Toc35539745)

# Background and Questions

Our team conducted various analyses of NFL offensive statistics from 1970 through 2019, which addressed the questions below: (*Please assume: the years from 1970-2019 were used to answer a question unless otherwise noted.*)

1. What were the different eras of the NFL and how many eras were there?
2. After identifying the clusters, does a modern era exist and if so, what years comprise the modern era?
3. What were the average yards gained (within eras as well)?
   1. Did passing or rushing yards contribute more to the yards gained and did it change over time (within eras as well)?
4. Did turnovers increase or decrease?
5. How many points were scored per season?
6. Did point differentials change (within eras as well)?
   1. What teams had the biggest point differentials in the regular season?
   2. What teams had the biggest point differentials in the Super Bowl?
7. Were more points scored early or late in the season (2019 season only)?
8. What teams were outliers in terms of total yards, rushing yards, and passing yards (modern era only)
   1. Did this translate to success?
   2. How did the outliers compare to Super Bowl teams, if the outlier did not win the Super Bowl?
9. In which playoff games did teams have exceptional performances in terms of total, passing, and rushing yards?
   1. Did this translate to success?

# The Data

Our team used three datasets, which were obtained from Pro Football Refence (https://www.pro-football-reference.com/). The website contains football data from 1920 through 2019, with more data added as time progresses. As an example, the 1920 dataset only contained offensive data related to the total points scored and number of touchdowns for each team whereas the 2019 contained those data and 23 additional variables for each team.

We used a subset of the data available by extracting the years 1970-2019. We selected 1970 due to it being the year the NFL merged the AFL and NFL. From 1960 through 1965 the two leagues were separated and did not play games against each other, which does not make for a fair comparison to the current NFL, in part, because teams now play inter-conference games. The leagues came together in 1966 to play the first Super Bowl, but inter-conference games still did not take place during the regular season. The leagues eventually merged in 1970, and at this point served as the beginning point of the NFL and our dataset. The start of the 1970 season is commonly referred to as The Merger. The data for each season was saved to three csv-files called “NFL\_SeasonResults”, “SeasonSummary\_1970\_2019”, and ”Modern”.

## ***NFL\_SeasonResults***

This dataset containes 26,369 games played during the regular season and postseason with 23 variables provided for each game. The dataset was used, in part, to answer questions 3-9. Due to the data not having to be sourced through Python we leveraged output of previous scrapings from R, and from data that was manually copied from the site.

## ***SeasonSummary\_1970\_2019***

This dataset contains 51 variables for each season’s average by team per game. Since the data was gathered from 1970-2019 an adjustment was made to account for the fact that the NFL played 14 games from 1970-1977 and then 16 games from 1978-2019. The website normalized the different seasons durations by dividing the league averages by 14 and 16 depending on the season, which was listed as column total called Avg Tm/G. This dataset was leveraged to conduct the cluster analysis by using K-means and answers, in part, questions: 1-5, and 8.

## ***Modern***

After the clusters were established, a dataset containing 23 unique offensive and special-team variables for the years defined as the modern era was created. This dataset provided answers, in part question: 8.

# The Analyses

First, we identified three clusters within the NFL with a slight overlap between clusters. We used the elbow method to determine that three clusters provided us with the best fit to the data provided. Based on the output we determined that the following clusters were defined as:

* Cluster 0: 2011-2019 (this was used as the modern era)
* Cluster 1: 1979-2010
* Cluster 2: 1970-1978

It appeared that a 4th cluster could be provided but the visual revealed too much overlap:

|  |  |
| --- | --- |
| 3 Clusters | 4 Cluster |
| A screenshot of a cell phone  Description automatically generated | A screenshot of a cell phone  Description automatically generated |

To further see the differences between the eras we reviewed and compared the total yards, rushing yards and passing yards to each era. We learned that total yards and passing yards increased by each era, while rushing decreased. The charts below show the gradual increases and decrease.

|  |  |  |
| --- | --- | --- |
| Yards Gained by Era | Passing Yards Gained by Era | Rushing Yards Gained by Era |
|  | | |

Seeing this trend strongly suggested that passing yards contributed more to the total yards of offense from 2011-2019while the 1970-1978 seasons relied more on the running game and it was closely balanced in the 1979-2010 seasons.

Since total offense is the sum of passing and rushing, we knew that plotting a trend line of passing and rushing yards would help us identify which category contributed more to total yards. We could see the lines nearly touching in in the mid and late 70s and then separating quickly just before 1980 and then spiking again around 2010. Not only did it show how much more the passing game contributed to the offensive game, but it also reinforced the accuracy of the clustering.

To further understand the clusters, we wanted to see if additional passing yards also meant more turnovers due to the defense having more opportunities to intercept passes (assuming more passing yards were a result of more passing attempts). As we saw turnovers decrease and passing yards increase we theorized that points should increase. Based on the final graph we can see that such a correlation exists. In future analyses we would like to run some linear models and calculate a statistic such as R2 to further prove the theory.

|  |  |  |
| --- | --- | --- |
| Passing and Rushing Yards  1970-2019 | Turnover Totals 1970-2019 | Total Points 1970-2019 |
|  | | |

We saw further proof of the distinctions between the clusters when we plotted the line chart and graphed a bar chart from 1970 – 2019 for point differentials. This statistic was calculated by subtracting the losing team’s score from the winning team’s score and dividing by the number of winning teams to reach the average for each year.

There is a sharp rise in the early 70s, and another dramatic spike after a minor decline in the mid-70s. This would explain the height in the 1970-1978 cluster. The rest of the graph has dramatic spikes and declines but does not change as much by the time it reaches 2019. This is seen in the bar graph heights as well.

|  |  |
| --- | --- |
| Point Differential 1970-2019 | Point Differential by Era |
|  |  |

It was also interesting to see the point differentials of the Super Bowl. This was done by following the same formula above, but it was only applied to 50 Super Bowls. The first four Super Bowls were omitted as described in the data summaries above.

|  |  |
| --- | --- |
| Score Differential Super Bowl: Over 1970-2019 | Score Differential Super Bowl: Over Eras |
|  |  |

We then identified games in which a single team won by an extraordinarily-high number of points in any game from 1970-2019. When we looked at the data we were confused to see two teams with point differentials at nearly 60. We investigated the online source to confirm no errors were made in the transfer of the data and confirmed the Rams beat the Falcons 59-0 in 2004, and the Patriots beat the Titans 59-0 in 2009.

|  |
| --- |
| Winning Margins for All Teams: 1970-2019 |
|  |

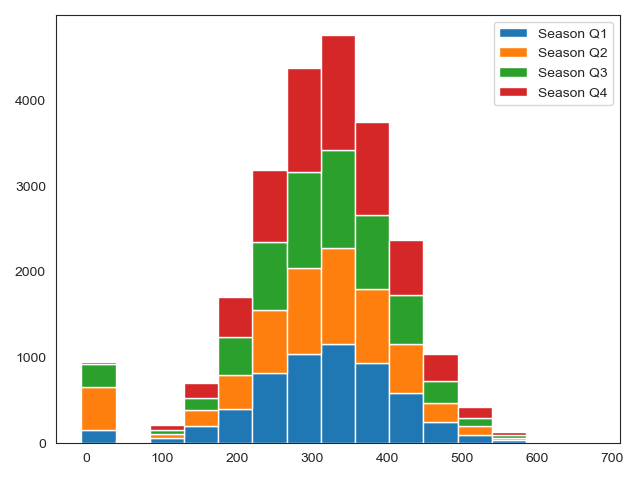
The Super Bowl box plot was interesting when you look at the 49ers. It shows they won a game by about 45 points, but it is not an outlier. Meanwhile a team like the Patriots has an outlier with 10 points. It is not entirely due to the Patriots having won six Super Bowls because the 49ers have five Super Bowls. What this shows is that the Patriots won several close Super Bowls while the 49ers won several blowouts.

|  |
| --- |
| Score Differential: Outlier Detection for Margin of Victory by Super Bowl Winning Teams: 1970-2019 |
|  |

As we focused on points scored we wanted to explore the distribution of points scored during specific times of the season. A potential theory would be that points scored declines as the season progresses due to injuries, fatigue, cold weather, and the added pressure of making the playoffs. In order to review this, we broke the NFL season into quarters which were defined as:

|  |  |
| --- | --- |
| Weeks (18 weeks in 1993 only) | Quarters |
| 1-4 | 1 |
| 5-8 | 2 |
| 9-12 | 3 |
| 13-18 | 4 |

What we found was a normal distribution of points scored and a stacked bar that showed points were evenly scored throughout the course of a season. The bars to the far left were counted as NAs, which was due to bye weeks included in the data. This added stacked bar was good to maintain as it proves we did account for bye weeks. Even further one can see the number of byes is extremely rare late in the season, which aligns with the data.



|  |
| --- |
| Points scored by grouped weeks: 2019 |
|  |

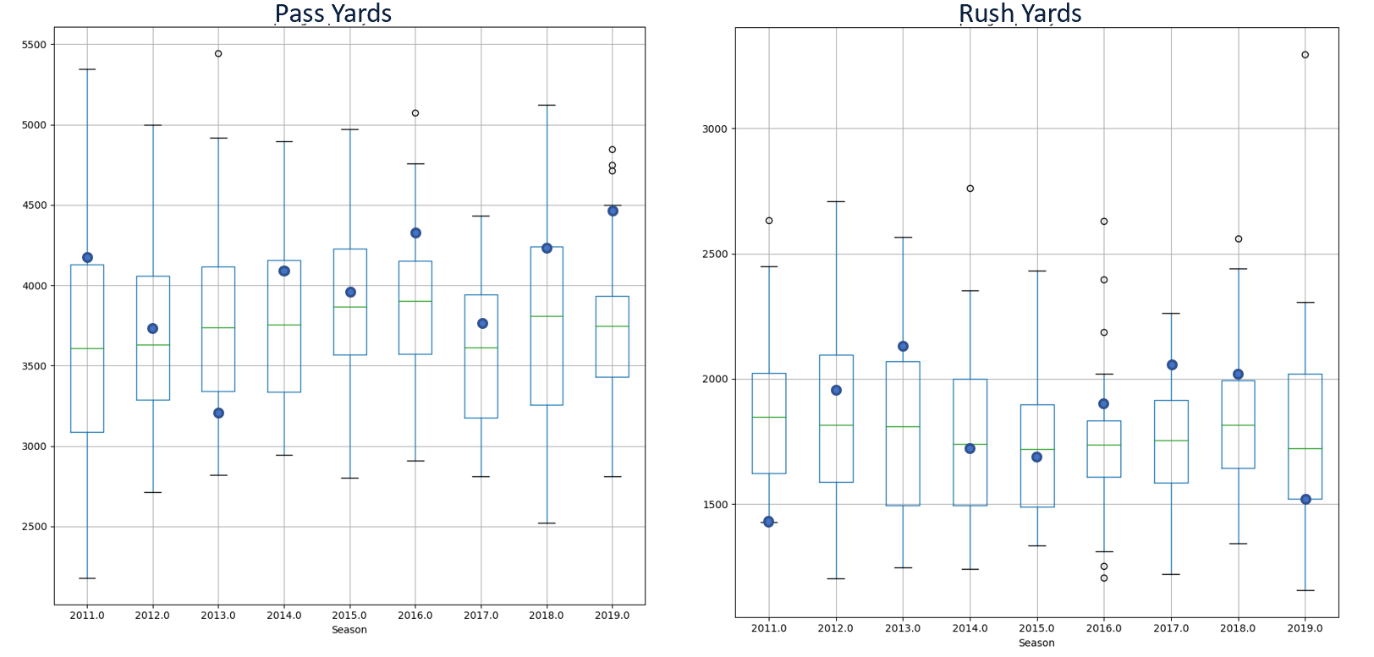
We also wanted to explore potential outliers during the modern era to see if exceptionally high totals in passing or rushing yards during the regular season led to success. Even though passing became more prolific we saw that four of the five teams who had exceptionally high passing yards in a season missed the playoffs while the other outlier lost in the Super Bowl. Meanwhile, the nine outliers for rushing yards made the playoffs five times with one team, the Seattle Seahawks in 2014, making the Super Bowl, but losing. Interestingly, Seattle lost the Super Bowl against the Patriots due to throwing an interception from the 1-yard line late in the fourth quarter instead of running the ball.

|  |  |
| --- | --- |
| Regular Season Passing Yard Outliers and Season Result: 2011-2019 | Regular Season Rushing Yard Outliers and Season Result: 2011-2019 |
| A picture containing sitting, white, water, light  Description automatically generated | |

When Super Bowl winners were plotted it was insightful to see that they won by having a much more balanced approach and not being overly dependent of one area of their game. The closest outlier was this past season (2019) with the Kansas City Chiefs passing game.

The table below lists the Super Bowl winners and the season in which they won. The boxplot shows where those teams are plotted for passing and rushing.

|  |  |
| --- | --- |
| Super Bowl Winner | Year |
| Giants | 2011 |
| Ravens | 2012 |
| Seahawks | 2013 |
| Patriots | 2014 |
| Broncos | 2015 |
| Patriots | 2016 |
| Eagles | 2017 |
| Patriots | 2018 |
| Chiefs | 2019 |



|  |  |
| --- | --- |
| Winning Super Bowl Team Regular Season Passing Yards (Team plotted with blue dot) | Winning Super Bowl Team Regular Season Rushing Yards (Team plotted with blue dot) |
|  | |

And finally, we identified teams that had exceptionally high rushing and passing yards in one playoff game. We detected a similar trend to that of regular season outliers in the sense that success was correlated more to the running game then the passing game. There are a couple of reasons as to why this is happening but would require a deeper analysis. Some of the theories would include:

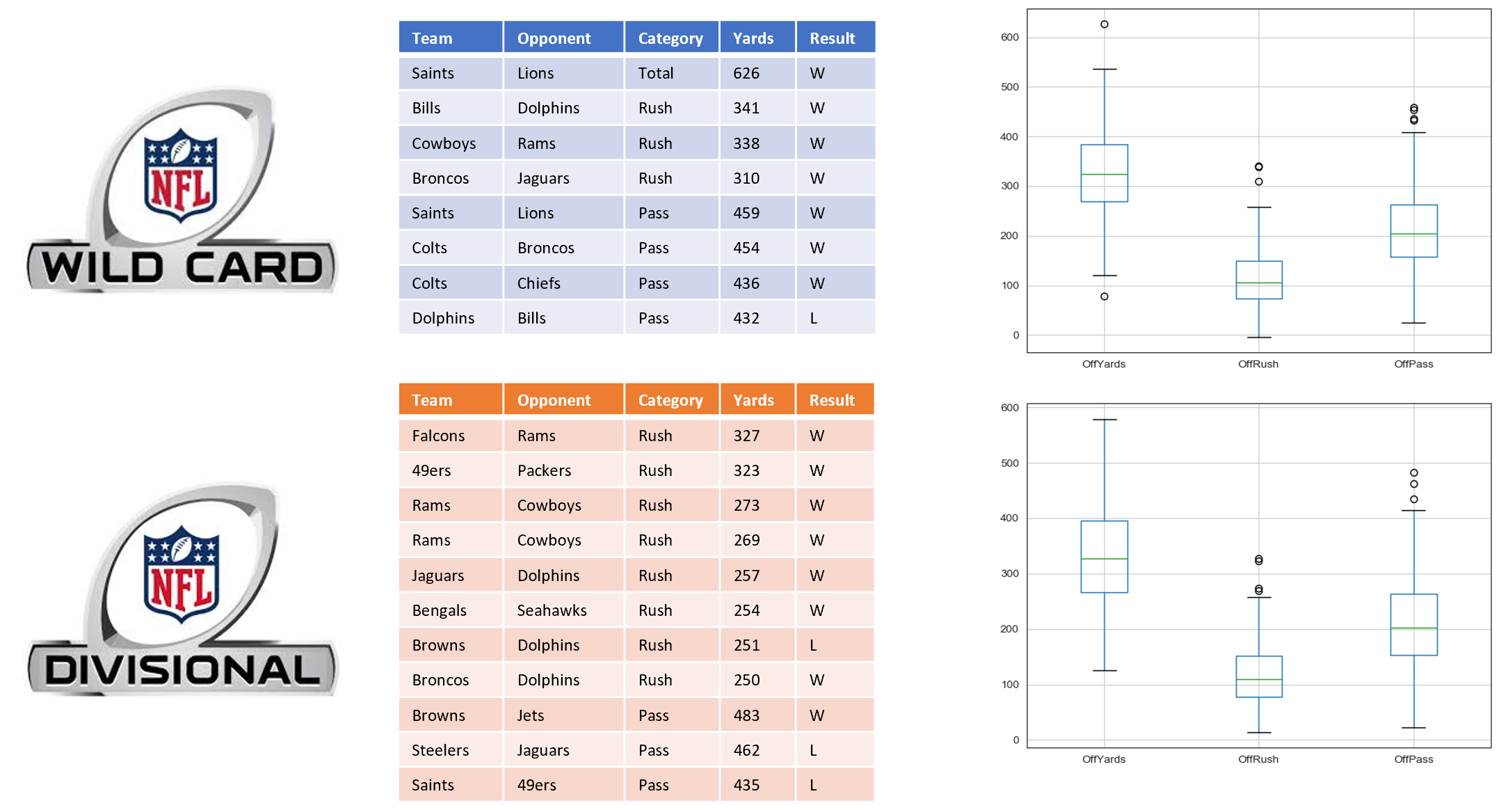
* Rushing yards may have lower variance. This means if a team is able to run the ball successfully, they are consistently gaining 3-5 yards per attempt whereas passing yards can be gained in bigger chunks, which means yards can be compiled quickly but not consistently.
* An effective rushing game wears down a defense and in turn makes it easier to score points in the long-term.
* An effective run game keeps the opponent’s offense off the field because time of possession increases, which leads to fewer opportunities for the opponent to score

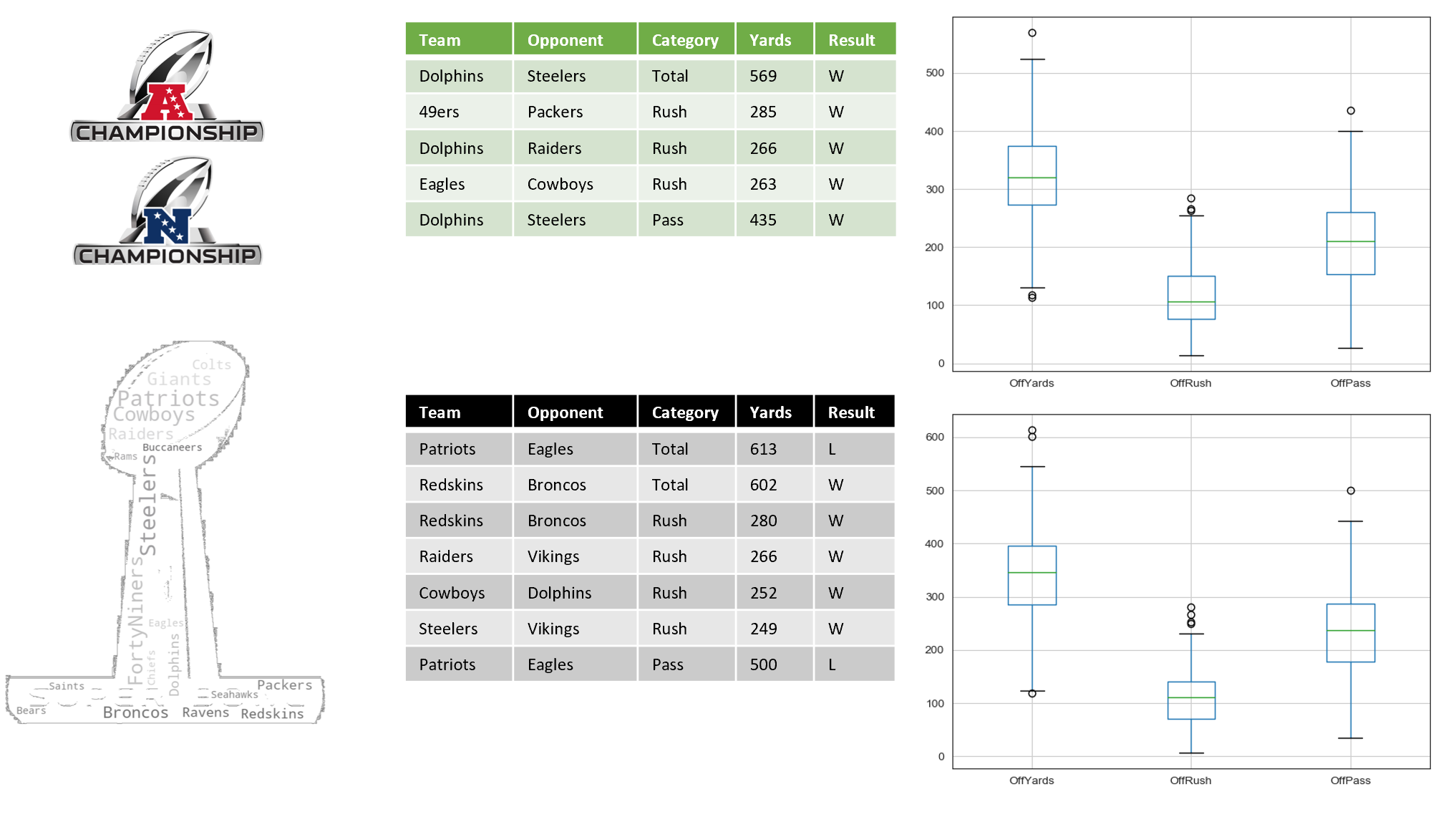
Regardless of the theory, teams that had exceptionally high rushing yards in playoff games were 17-1 in the postseason, including 4-0 in the Super Bowl. Meanwhile, teams with exceptionally high passing yards were just 5-4 in the postseason and 0-1 in the Super Bowl.

Outliers and Result

Boxplot Outlier Detection

Round





Final Thoughts and Next Steps

It was clear to see the increase of passing yards and its impact on scoring from 1970-2019 and within the different eras but being an outlier within passing during the modern era and post-season did not provide any benefits. In fact, it appeared that being an outlier in the running game provided more benefit in terms of making and then winning in the playoffs.

The boxplots with Super Bowl winners also revealed that the best approach may be to have a balanced offense in terms of passing and rushing yards, which produces more wins due to the offense being less predictable and more difficult for defenses to defend. And despite points increasing over time there are clearly dips in the chart, which suggests that defense learn how defend new game plans, but it does not last for prolonged periods of time.

Further analyses would be needed to identify the factors that lead to wins during the regular season and postseason such as multiple regression or building probability models through machine learning. A predictive model could also be built that runs throughout the season and assigns the probability of a team winning the Super Bowl.