Katie Livingston, Giovanni Macaluso, Rocco Patel
OA3801 Final Project

Executive Summary

In the 2023 football pre-season, controversy emerged when a group of National Football League (NFL) running backs questioned the merits of the pay system for each football position. Specifically, an argument was raised on whether running backs are being appropriately valued based on their contributions to team wins. With the NFL season now in full swing, the goal of this research was to find supporting evidence on whether running backs are really worth the money they are asking for.

The research setting consisted of all 32 NFL teams over the span of the last 10 years (2013-2022). Regardless of team location or name changes throughout this time period, the data was compiled for each overall franchise. The primary data sources included Pro Football Reference and Over the Cap. Pro Football Reference is a robust source for current and historical NFL players, teams, and leaders. Within this site, we downloaded individual Excel Workbooks for data at the team and individual player level. For each of the 32 teams and for each year within the last 10 seasons, we collected data on rushing and receiving yards for every player. Along with metrics on rushing, receiving and total yards, these workbooks also provided data on player age, position, and games played and team wins and losses. Furthermore, the second part of the equation was to compile data for the overall team and positional salary amounts for each of the 32 teams over the last 10 seasons. This information was obtained from Over the Cap, which is a website that tracks NFL salary information, including salary caps, player contracts, and salaries by position.

After identifying the problem and sources of data with which to evaluate this problem, we were able to begin the data analysis workflow. The first step was the data curation. We obtained over 350 individual Excel Workbooks; however, the true computational power comes when looking at the information in the aggregate. Within the JupyterLab and Python coding environment, we utilized Os scripting, Glob, and Openpyxl to read in each workbook to the environment. Next, we parsed through each dataset to clean up any missing values or values which may be different across the data sources. Finally, all the data was merged into a Pandas dataframe to view and work with data for all 32 teams over the last 10 seasons.

Throughout this data curation process, a few challenges emerged. For each of the Excel Workbooks, the first two rows both contained header information and some of the columns were named the same value. This presented an issue when creating the data frame because the first header row needed to be skipped and some of the names of columns in the second row needed to be changed to more descriptive and unique values. Furthermore, the Excel Workbooks used an older file extension format, .xls, which was not compatible with the package Openpyxl. Therefore, we put each workbook through a file converter system found online to convert them all to the newer .xlsx format. When parsing the multiple sources of data, we also found some cases where the team name values were written as “City Mascot” rather than just “Mascot”. For such cases, we made each team name just the mascot name. The salary data presented its own challenges, for the data could not be readily downloaded in a specific format from the website.

Instead, we copied the salary information for the past 10 years into a new Excel file with a tab for each year and added a year column to the data. The addition of the year column allowed us to merge the tabs together in Pandas and concatenate with the other sources of player and team data.

The second step in the data analysis workflow was computation. The overall goal was to compare player production and team success with how much the position and its members are getting paid. Working within the main Pandas dataframe, we utilized GroupBy operations to explore relationships across a variety of fields in the data. Rather than begin the project with the assumption that the NFL running backs are correct that they are undervalued and find evidence to support this claim, we used the data to analyze and visualize the situation as a whole. To analyze the data, we utilized RStudio to run linear regression models. We ran a model of all the individual offensive positions as well as the total defensive spending against wins to see if any of them were statistically significant in predicting wins. The resulting equation for predicting the number of wins based on the dollar amounts spent on each category can be seen below in Equation 1. The equation says that for every dollar spent on each position then the associated increase in wins would be the assigned coefficient for that position.

$$\begin{aligned}\widehat{Wins} = 4.38 &+ (6.32 * 10^{-8}) * RB \\ &+ (1.12 * 10^{-8}) * WR \\ &+ (6.74 * 10^{-8}) * TE \\ &+ (4.50 * 10^{-8}) * QB \\ &+ (-2.04 * 10^{-8}) * OL \\ &+ (3.54 * 10^{-8}) * Defense\end{aligned}$$

Equation 1: Wins vs Salary Variables

After running the linear regression model, we found that only two of the variables were statistically significant. The first variable was quarterback salaries which had a p-value of .02453, which is significant at the .05 level. The second variable was total defense salary which had a p-value of .00287, which is significant at the .005 level. However, running back salary was not shown to be statistically significant at any level with a p-value of .20291.

The primary workflow for the project was centered around procuring and processing the data into a usable format for computation and analysis. The end state of this workflow was to explore new ways to visualize the data and find trends that may help answer the initial problem. We utilized the Python packages Matplotlib and Plotly to display static and interactive graphs and visuals. A variety of visuals were produced; however, the most significant relationships were found in the following visuals.

Looking across all 32 NFL teams over the past 10 seasons, we conclude that spending more on running backs does not produce more team wins. Teams that have made it to the playoffs or even won the super bowl have spent no more, and in many cases less, money on running backs than teams that have performed much worse in a season. Additionally, while NFL teams have a significant amount of revenue, each team still has to work within the confines of their own salary caps. A tradeoff is required to balance the salary demands of every skilled football position. A "star" player in another position such as quarterback has a much greater correlation to team wins and is thus paid more. The depth of running backs on a team, rather than one flashy player, has much greater value to the team.

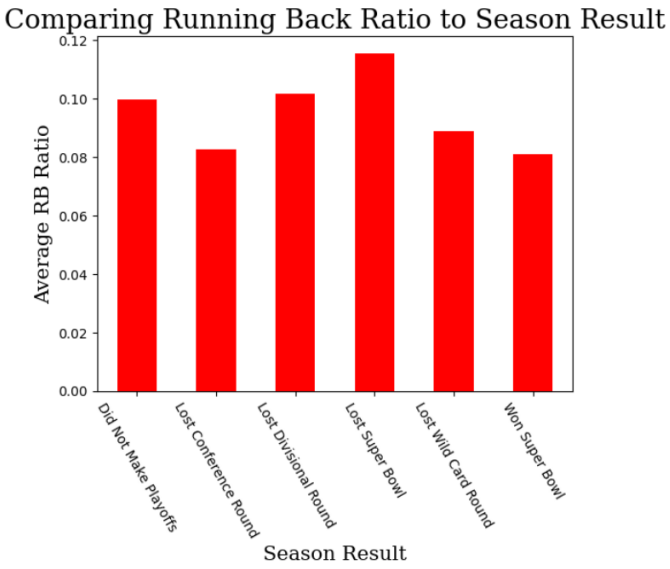

Figure 1: Comparing average spending on running backs to season result.

In conclusion, we found no significant indication that running backs should be paid more based on their contributions to team wins. For future research into this topic, we may consider a deeper analysis on whether defense or offense is more important for winning championships and exploring the data greater at the individual player level. For the data analysis workflow process, we may improve this project through the use of additional computational methods and enhanced visualizations. This includes more efficient data collection through fuzzy matching, web scraping, and API calls. Additionally, we could use plots with customized icons and put all the visuals into a dashboard.

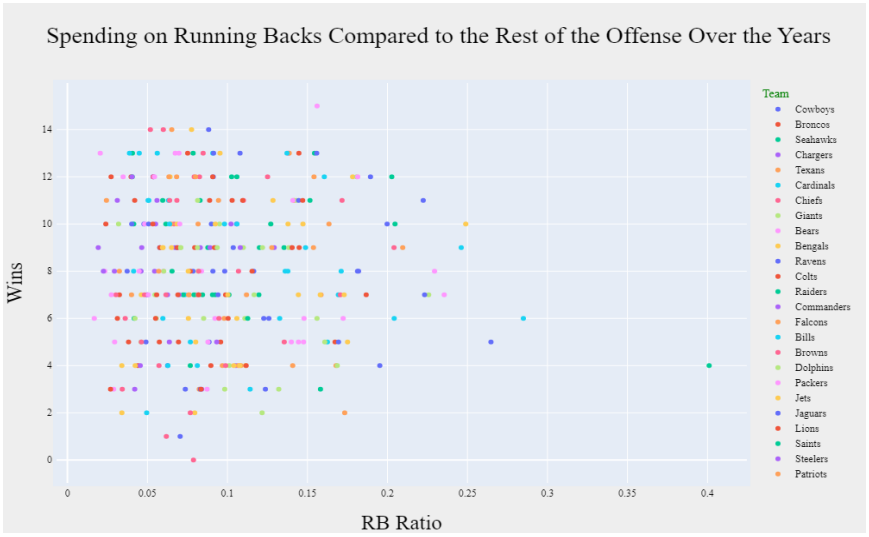

Figure 2: Wins compared to RB Ratio, broken down by team

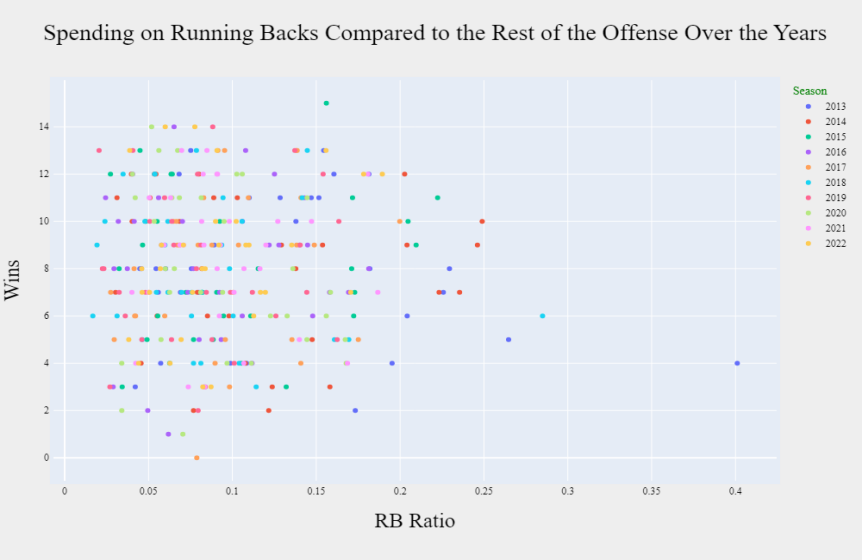

Figure 3: Wins compared to RB Ratio, broken down by year