Jake Aufiero

Professor Nathan Carter

MA346

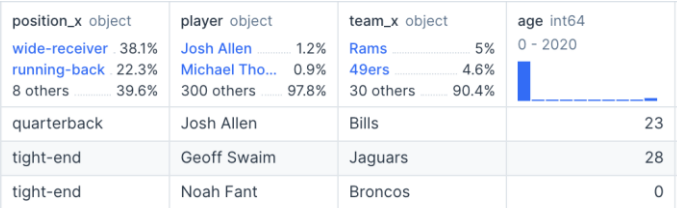
19 Dec 2023

Beyond the Goalposts: Statistical Modeling of NFL Running Backs

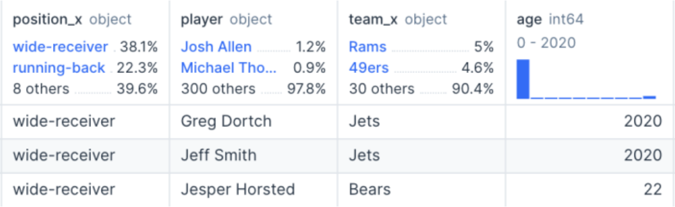
Commonly regarded as one of the most popular sports worldwide, football maintains a strong influence on the sporting world, uniting cultures and providing opportunities for individuals of all backgrounds. The NFL alone brought in $18.6 billion in revenue in 2022 and represents one of the largest known federations in sports. Player valuation, in particular for running backs, has become increasingly relevant in recent years, leading teams and individuals to pose the question of what factors most significantly impact how much a player is currently worth. I will begin by introducing the datasets I worked with, explaining how I chose to merge them and any cleaning that was required. Then, I'll provide an exploratory analysis on the data itself, highlighting any noticeable trends and/or features within it. Next, I'll aim to answer the aforementioned question utilizing regression analysis and hypothesis testing. By the end of the report, I will show that the total amount of yards a player accumulates, the number of carries, and the number of snaps played throughout the season predict around 72% of the variability, in total, of the value of an NFL running back's contract, and are three of the most significant variables in predicting how much a player is worth.

The two datasets I used for this project were both found on Kaggle through a manual search. The first dataset, which I refer to as "df1" in my code, contains up to date salary information on all offensive and defensive players in the NFL. This includes the total value of the player's contract, the amount of money they receive per year, how much of it is guaranteed, and when they will become a free agent, alongside other standard information. The second dataset, understandably named "df2" in my code, contains relevant statistics for all offensive skill position players and a few special teams players from the 2012-2022 NFL seasons. Examples of some of the variables in this dataset are the number of touchdowns the player scored, number of yards accumulated, and the number of snaps played throughout the season. Since my question is about how much a player is currently worth, I only required the data from the most recent season (2022/23), so I reduced this dataset down to only include the information I desired. These datasets fit my goals as I want to see how a player's current value (from df1) can be affected by their in-game statistics (from df2). So, merging them together allows me to perform a statistical analysis on one joint dataset that provides insight into the key factors for determining a player's worth.

In order to get the data into a suitable state, I first merged the salary dataset with the reduced statistics dataset for the 2022 season on the players' names using an inner join, meaning I am creating a new, cohesive dataset that contains information for players whose names matched across the two individual datasets. Because the statistics dataset consisted of primarily offensive players, my merged dataset contains only offensive skill position players and a couple of special teams players. This is acceptable as my main focus is on running backs, who are part of the offense. Now, the merged dataset needed a little bit of cleaning before I was able to analyze it. First, a few players' ages were incorrect, either being 0 or 2020. Examples of this are below:



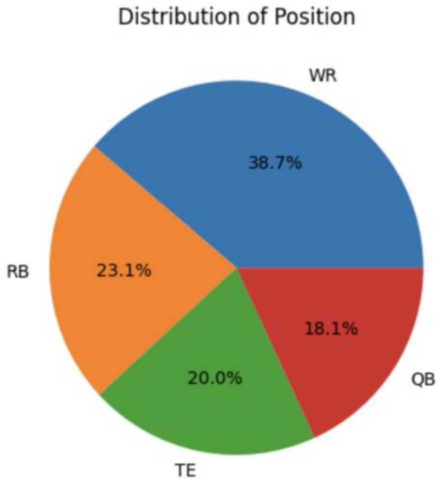
position_x object	player object	team_x object	age int64
wide-receiver 38.1% running-back 22.3% 8 others 39.6%	Josh Allen 1.2% Michael Tho... 0.9% 300 others 97.8%	Rams 5% 49ers 4.6% 30 others 90.4%	0 - 2020
quarterback	Josh Allen	Bills	23
tight-end	Geoff Swaim	Jaguars	28
tight-end	Noah Fant	Broncos	0



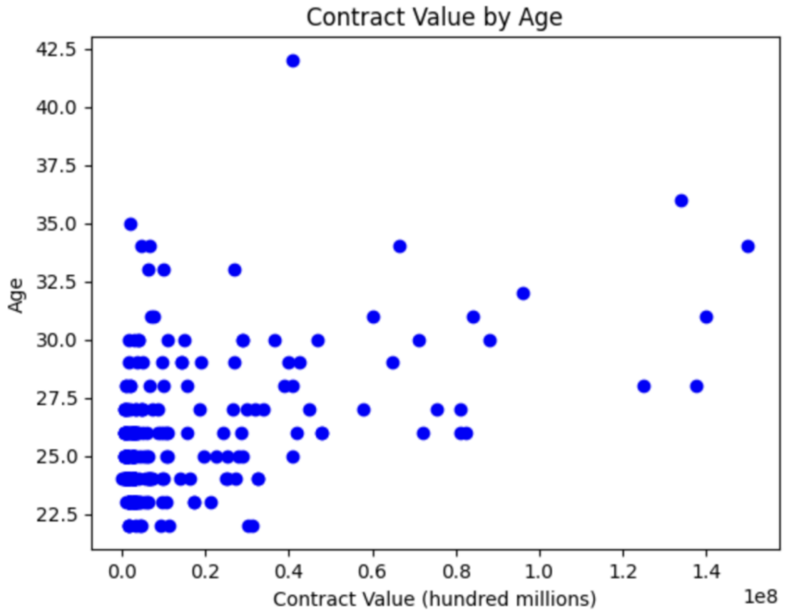
position_x object	player object	team_x object	age int64
wide-receiver 38.1% running-back 22.3% 8 others 39.6%	Josh Allen 1.2% Michael Tho... 0.9% 300 others 97.8%	Rams 5% 49ers 4.6% 30 others 90.4%	0 - 2020
wide-receiver	Greg Dortch	Jets	2020
wide-receiver	Jeff Smith	Jets	2020
wide-receiver	Jesper Horsted	Bears	22

I replaced these values with numPy's missing value indicator "np.nan", so that a statistical analysis would avoid thinking these players were 0 or 2020 years old, and thus fail to affect the outcome of the analysis. Next, I removed some irrelevant columns, such as *season_type* (was "REG" for every entry, even though it is known that the data is from the regular season), *fantasy_points*, and *fantasy_points*_ppr, the latter two having to do with fantasy football, which is irrelevant in the context of this report. After that, I noticed that one player, Josh Allen of the Buffalo Bills, was repeated four times throughout the data. So, I removed three of these rows so that Josh represented one row and his repeated stats and contract information would not skew the analysis. Finally, I decided to rename the position and team columns so that there would be no confusion as to what they represent for an onlooker unfamiliar with the data. With the cleaning completed, I can now provide an introductory analysis of the data.

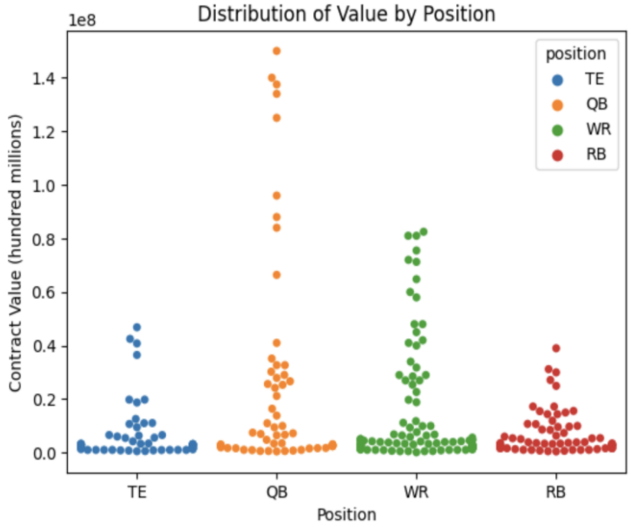
Offensive skill position players include quarterbacks (QB), running backs (RB), wide receivers (WR), and tight ends (TE). Out of these positions, I first wanted to see what proportion was represented by running backs.

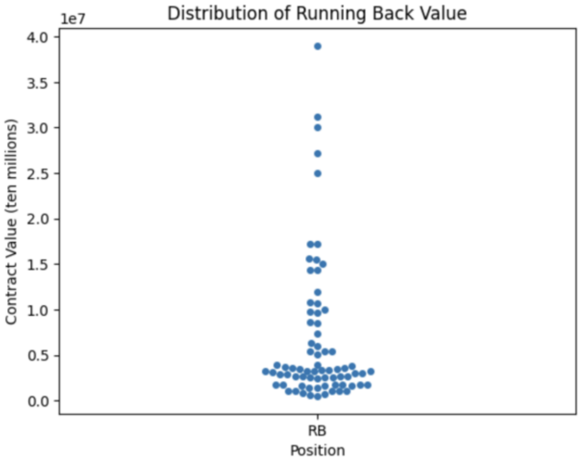


As depicted above, running backs make up 23.1% of offensive skill position players, coming second only to wide receivers in terms of quantity. One factor I believe could affect how much a player is worth is age. As a player gets older, it's common for their contract to be smaller, although this isn't always the case. The visualization of this idea is below:



There appears to be a slightly positive linear relationship between the age of a player and how much money their contract is for. This was a little surprising to see, yet counterintuitively explained by the fact that rookie contracts (signed when a player is drafted) are typically smaller than an average NFL contract as the player is just beginning their journey in the league. I then wanted to see the distribution of how each individual position is paid, as well as how running backs are paid.





Quarterbacks appear to be worth the most out of the four skill positions, followed by wide receivers. Running backs and tight ends seem to be relatively equal in terms of contract value. This makes sense as quarterbacks are commonly thought to be the most valuable players in the league and are typically the public faces of their respective teams. The running back position does not get much attention when it comes to contract discussions, which is another reason why I want to explore which factors most significantly affect their value.

My first thought was to see how the number of yards a running back accumulates throughout the season affects their value. At the surface level, it seems likely that a player with a larger number of yards would be worth more than another player with less yards. In order to check if this variable is significant in predicting contract value, we can utilize simple linear regression. The dependent (response) variable will be *total_value* and the independent (predictor) variable will be *total_yards*. This independent variable encompasses all rushing and receiving yards so it will do a better job of ironing out the differences between running backs who only run and those who also catch passes. Based on the data for running backs, the regression equation is as follows (y represents *total_value)*:

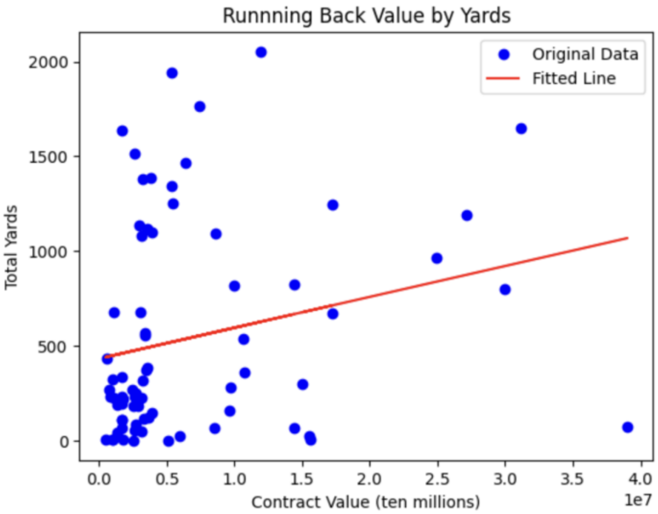
$$y = 0.0000163 + 433.02*\textit{total\_yards}$$

This means that a one-yard increase in the number of yards accumulated increases the value of a running back by $433.02. In order to test whether *total_yards* is significant in predicting *total_value*, we can conduct a hypothesis test (let $\alpha = 0.05$):

$$H_0: \beta_{\text{total\_yards}} = 0$$

$$H_a: \beta_{\text{total\_yards}} \neq 0$$

The individual p-value from the regression is 0.048. At a confidence level of 0.05, we can see that the p-value < α, leading us to reject H_0 and determine that *total_yards* is significant! The R^2 value for the test was 0.23, meaning that *total_yards* can predict around 23% of the variability in an NFL running back's contract. The chart below shows the data from the regression.



The next two variables that I thought could significantly predict a running back's value were the number of carries they had thoughout the season and the total number of touchdowns they scored. When conducting a regression for each of these variables, I found their respective equations:

$$y = 0.000003 + 70.43*\textit{carries}$$

$$y = 0 + 2.94*\textit{total\_tds}$$

The individual p-values for *carries* and *total_tds* were 0.049 and 0.059, respectively. At a confidence level of 0.05, *carries* appears to be significant, but *total_tds* is insignificant. So, we need to see what other variables might better predict value. When it comes to *carries*, a one-carry increase for a running back will cause their value to increase by $70.43. The R^2 value for *carries* was 0.23, meaning that this variable can also predict around 23% of the variability in a running back's contract.

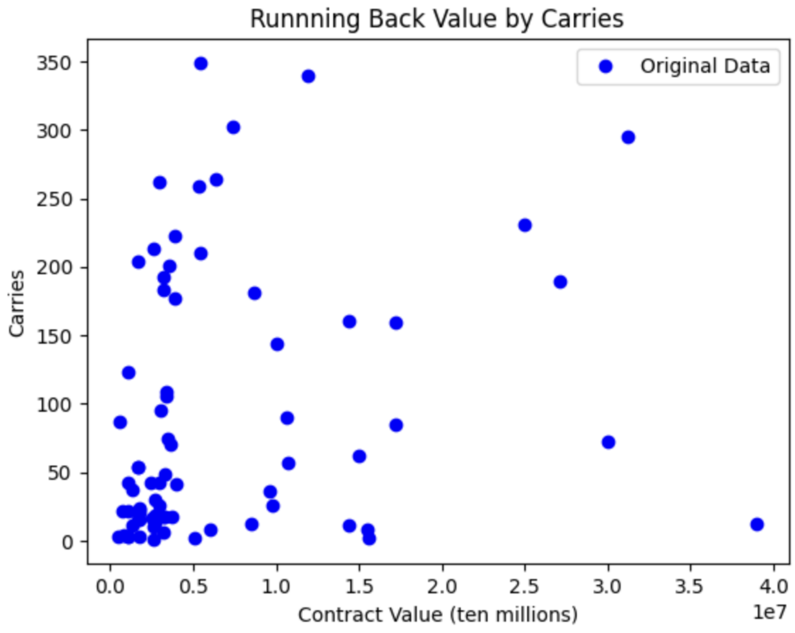
The final concept that I believed could affect how much a player is worth is how healthy they are. Now, there wasn't a variable in the dataset to represent a running back's health, but there were some that could correlate to it. For example, *offense_snaps* represents the number of plays the player was involved in throughout the season, which essentially shows how often they were playing and how healthy they were. When running a regression on this variable, the equation is as follows:

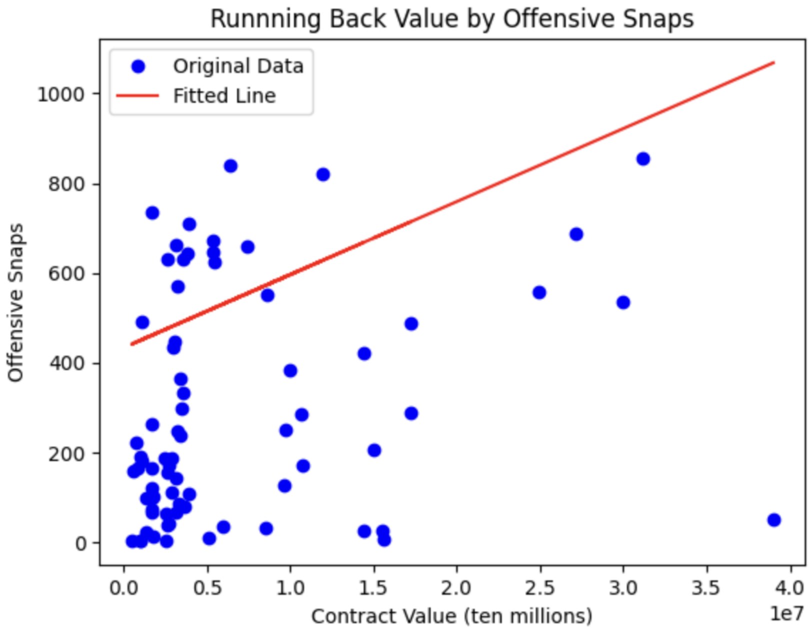
$$y = 0.000008 + 229.28 * \textit{offense\_snaps}$$

The individual p-value for this variable is 0.026, so this *offense_snaps* is significant at a 0.05 confidence level. The R^2 value is 0.258, so it can predict another 25.8% of variability in the dependent variable. Similarly, a one-snap increase would result in an increase to the player's value of $229.28. Please refer to the appendix for the graphs of *carries and offensive_snaps*.

In conclusion, the total number of yards, number of carries, and number of offensive snaps played throughout the season are three of the most significant variables in predicting the contract value of an NFL running back. Together, they predict a total of 71.8% of the variability in contract value and can give teams and executives insight into how much a player they have on their roster or are looking to sign is truly worth. Playing the greatest amount of snaps as possible gives running backs the best chance to increase their value as they are likely to generate more yards and carries by being on the field and healthy. In regard to future work, I would like to determine if there is any discrimination between the way that offensive and defensive players are paid, but that study is ultimately for another day.

Appendix:





Works Cited

Antonov, Aleksandr. “Football Players’ Salaries.” *Kaggle*, 6 June 2019, www.kaggle.com/datasets/trolukovich/football-players-salaries/.

Hyde, Philip. “NFL Stats 2012-2022.” *Kaggle*, 6 Aug. 2023, www.kaggle.com/datasets/philiphyde1/nfl-stats-1999 2022select=yearly_data_updated_08_23.csv.