**NFL Player Value Analysis (2022–2024)**

**1. Introduction**

The NFL is one of the most competitive and financially intense leagues in sports. With salary caps and massive contracts, teams must make smart, data-driven decisions to get the most out of their rosters. In this project, I wanted to find out whether NFL players are actually paid in line with how they perform on the field. I focused on quarterbacks (QBs), running backs (RBs), and wide receivers (WRs) across the 2022 to 2024 seasons.

To do this, I combined player salary and performance data from two well-known football websites. I used scraping and data wrangling to clean and organize the information, then created normalized stats to fairly compare players across positions and years. My analysis includes visualizations, a statistical test comparing positions, and machine learning models to see how well performance can predict salary. In the end, the goal is to better understand which players — and which positions — give teams the most value for their money.

**2. Data**

**2.1 Data Sources and Scraping**

For this project, I pulled data from two public football websites. I scraped performance stats for quarterbacks, running backs, and wide receivers from Pro Football Reference for the 2022 to 2024 seasons. I also got salary information for the top 35 players at each position from OverTheCap for the same years.

To collect the data, I used pandas.read\_html() to grab tables from Pro Football Reference, and BeautifulSoup to scrape the salary tables from OverTheCap. After that, I matched players from both sources using their names and the year, and cleaned things up by removing duplicates, filling in missing values, and converting everything into usable formats.

**2.2 Data Cleaning and Merging**

After scraping the data, I cleaned the datasets by getting rid of extra formatting (like dollar signs and commas), converting columns to numbers, and dealing with missing values. I only kept players who had both salary and performance data available. If a stat didn’t apply to a player’s position — like rushing yards for QBs — I filled it in with zero. I also removed any duplicate headers and empty player rows. The final merged dataset, called merged\_df, included all the key columns I needed for the analysis.2.3 Data Dictionary

**2.3 Data Dictionary**

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Type | Source | Description |
| Player | Text | Both | |  | | --- | |  |  |  | | --- | | Full name of the player | |
| Year | Numeric | Both | |  | | --- | | NFL season year (2022, 2023, or 2024) |  |  | | --- | |  | |
| Team | Text | Pro Football Reference | Team abbreviation (e.g., KC, BUF) |
| Position | Text | Both | Player position (QB, RB, or WR) |
| Salary | Numeric | OverTheCap | Annual salary in USD |
| Passing\_Yards | Numeric | |  | | --- | | Pro Football Reference |  |  | | --- | |  | | Total passing yards (QBs only) |
| Pass\_TDs | Numeric | |  | | --- | | Pro Football Reference |  |  | | --- | |  | | |  | | --- | |  |  |  | | --- | | Total passing touchdowns (QBs only) | |
| Completion\_Pct | Numeric | |  | | --- | | Pro Football Reference |  |  | | --- | |  | | Pass completion percentage (QBs only) |
| Rush\_Yards | Numeric | |  | | --- | | Pro Football Reference |  |  | | --- | |  | | Total rushing yards (RBs only) |
| Rush\_TDs | Numeric | |  | | --- | | Pro Football Reference |  |  | | --- | |  | | Total rushing touchdowns (RBs only) |
| Yards\_per\_Carry | Numeric | |  | | --- | | Pro Football Reference |  |  | | --- | |  | | Average yards per carry (RBs only) |
| Receiving\_Yards | Numeric | |  | | --- | | Pro Football Reference |  |  | | --- | |  | | Total receiving yards (WRs only) |
| Receiving\_TDs | Numeric | |  | | --- | | Pro Football Reference |  |  | | --- | |  | | total receiving touchdowns (WRs only) |
| Yards\_per\_Catch | Numeric | |  | | --- | | Pro Football Reference |  |  | | --- | |  | | Average yards per reception (WRs only) |
| Stat\_Total | Numeric | |  | | --- | | Calculated |  |  | | --- | |  | | |  | | --- | |  |  |  | | --- | | Total production = yards + (TDs × 20) | |
| Normalized\_Stat | Numeric | |  | | --- | |  |  |  | | --- | |  | |  |  |  | | --- | | Calculated | | Min-max scaled Stat\_Total |
| Normalized\_Salary | Numeric | |  | | --- | |  |  |  | | --- | | Calculated | | Min-max scaled Salary |
| Value\_Index | Numeric | |  | | --- | |  |  |  | | --- | | Calculated | | Normalized\_Stat – Normalized\_Salary |
| Normalized\_Yards\_per\_Million | Numeric | |  | | --- | |  |  |  | | --- | | Calculated | | Normalized\_Stat / Normalized\_Salary |
| Player\_Year | Text | |  | | --- | |  |  |  | | --- | | Calculated | | Player name with year (e.g., “Joe Burrow (2023)”) |
| Position\_RB | Binary | |  | | --- | |  |  |  | | --- | | Calculated | | 1 if player is RB; 0 otherwise (used in ML) |
| Position\_WR | Binary | |  | | --- | |  |  |  | | --- | | Calculated | | 1 if player is WR; 0 otherwise (used in ML) |

**3. Analysis**

**3.1 Value Metrics and Descriptive Analysis**

To evaluate how player performance relates to salary, I created a few custom metrics that helped compare players across positions and years. The first metric I used is called **Stat\_Total**, which combines all yardage categories (passing, rushing, receiving) along with touchdowns. I gave touchdowns extra weight since they have a bigger impact on games, multiplying each TD by 20. This helped balance out players who might not have the highest yardage but still score a lot.

Once I had a total performance score, I scaled both **Stat\_Total** and **Salary** using min-max normalization so everything would be on the same 0–1 scale. This was important because salary ranges can be drastically different between positions — for example, QBs tend to earn a lot more than RBs, which would make raw comparisons unfair.

From there, I created the **Value\_Index**, which subtracts a player’s normalized salary from their normalized stat total. A higher Value Index means the player performed better than expected based on their pay, while a lower value suggests they may be overpaid relative to their stats.

I also calculated a metric called **Normalized\_Yards\_per\_Million**, which divides normalized performance by normalized salary. This acts like an efficiency rating — showing how much production a team gets for each unit of salary. Together, these metrics let me identify which players and positions are delivering the most value for the money. The visualizations below show how this plays out across different roles and salary levels.

A green and white chart

AI-generated content may be incorrect.

Figure 1 - Top 10 player-seasons by Value Index

A graph showing the value of a graph

AI-generated content may be incorrect.

Figure 2 - Bottom 10 player-seasons by Value Index

A graph of different colored squares

AI-generated content may be incorrect.

Figure 3 - Normalized Yards per Salary Dollar by Position

A graph of a number of colored dots

AI-generated content may be incorrect.

Figure 4 - Normalized Stat vs. Normalized Salary (Scatterplot)

**3.2 Hypothesis Testing**

To better understand how value differs across positions, I conducted a Welch’s t-test comparing the average Value Index of quarterbacks (QBs) and running backs (RBs). This test was appropriate because it doesn’t assume equal variance between the two groups, which is important given the different salary structures and roles these positions tend to have.

The results showed a T-statistic of 3.71 and a p-value of 0.0002, which is well below the common significance threshold of 0.05. Based on this, I rejected the null hypothesis and concluded that there is a statistically significant difference in the average value between QBs and RBs.

These results support what I observed in the earlier visualizations: running backs tend to deliver more performance for their salary compared to quarterbacks. While this doesn’t mean that all RBs are underpaid or that all QBs are overpaid, it does suggest that, on average, RBs offer better value across the seasons analyzed.

**3.3 Machine Learning Models**

To see whether player performance could be used to predict salary, I built and tested three regression models: Linear Regression, Ridge Regression, and Lasso Regression. These models allowed me to compare the effect of regularization on prediction accuracy.

The input features I used included Stat\_Total (which combines yards and weighted touchdowns), Normalized\_Stat, the year of the season, and two dummy variables for position — one for wide receivers (WR) and one for running backs (RB), with quarterbacks (QB) serving as the baseline.

Before training the models, I standardized all the input features so they would be on the same scale. I then split the dataset into a training set and a test set using an 80/20 ratio. Model performance was evaluated using R², which measures how much variation in salary is explained by the model, and Mean Absolute Error (MAE), which gives the average error in predicted salary.

A graph with blue dots and a red line

AI-generated content may be incorrect.

Figure 5 - Predicted vs. Actual Salary – Linear Regression

**Results:**  
All three models performed similarly, each achieving an R² of approximately 0.22 and a Mean Absolute Error (MAE) of around $5.4 million. This suggests that while performance metrics offer some predictive power, they only explain a small portion of the variation in player salaries.

The scatterplot comparing predicted versus actual salaries reveals a general upward trend but also shows a widespread — particularly for players with higher salaries. This indicates that factors beyond on-field performance, such as contract structure, position market trends, or player reputation, likely play a significant role in determining salary outcomes.

**4. Conclusion**

This project examined how NFL player salaries relate to on-field performance by analyzing real data from the 2022 to 2024 seasons. By scraping and merging salary and performance statistics, I created normalized metrics that allowed for fair comparisons across positions and years.

The analysis revealed several key findings. Running backs consistently delivered the highest value relative to their salaries, and the difference between RBs and QBs was statistically significant. Additionally, multiple underpaid players were identified using the Value Index metric, highlighting how some players outperform their compensation. The machine learning models—Linear, Ridge, and Lasso—were able to explain about 22% of the variation in salaries, showing that while performance matters, it’s only part of the overall picture.

Looking ahead, future analyses could be improved by incorporating more context, such as player age, injury history, team performance, or contract structure. Overall, this project demonstrates how tools like web scraping, data cleaning, normalization, and modeling can be applied to uncover meaningful insights in the world of sports analytics.