Introduction

The problem we are tackling is to figure out if NFL players' salaries are overpaid or underpaid based on their performance. This problem is interesting because there are more than 1500 NFL players and some players are likely receiving a salary that is below or above their performance. Looking at how many teams or players that made a wrong decision by signing contracts that don't reflect the player's true value is interesting to us. Moreover, this is important and motivating because having a model to determine salary based on a player's performance can help both the team and player sign a fair contract. This can allow the team to make more reliable decisions and improve their spending habits. This can help protect the player's true value and potentially motivate the player to perform better.

Data

salary_proj_df:

We used two datasets here. First one is a CSV file called "salary_proj_df." This dataset contains salary-related data for NFL players. We accessed this dataset from the website called spotrac.com. The datasets contains useful columns such as:

Player: The full name of the NFL player.
Position: The playing position of the player, typically abbreviated (e.g., QB for Quarterback).
Base Salary: The annual base salary of the player in US dollars.

Specifically, as this dataset provides information on a player's salary, we will use the salary data in this dataset and compare it to the player's performance from our second dataset.

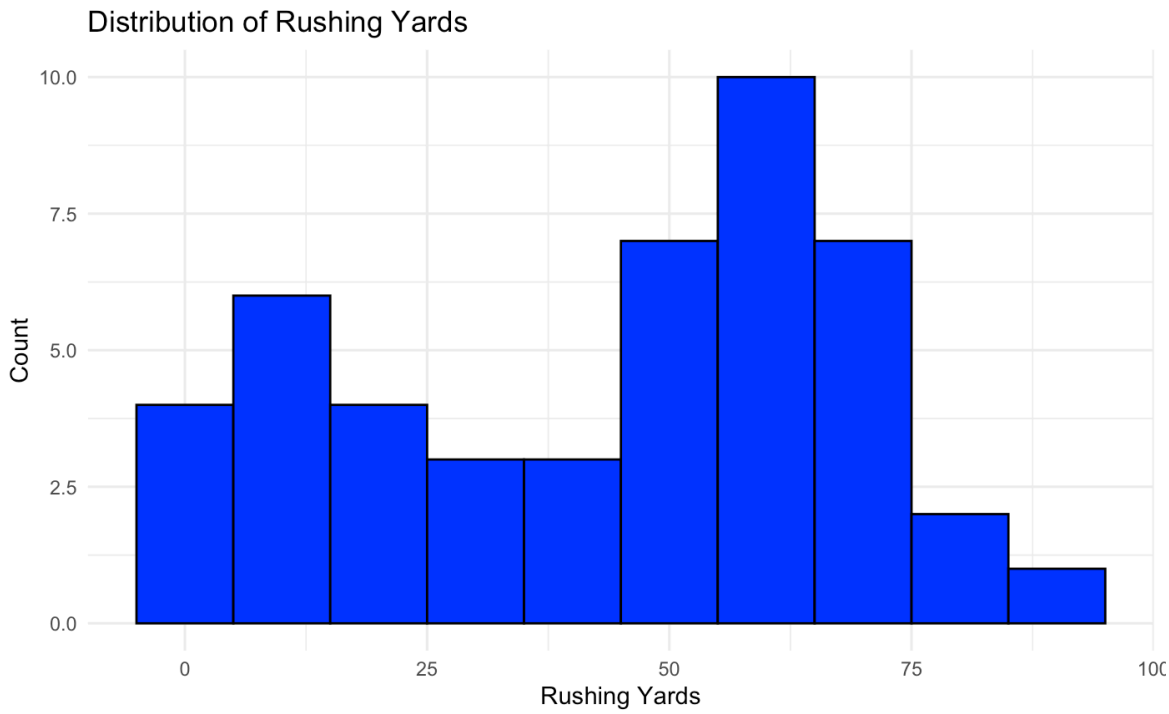
nflverse_df:

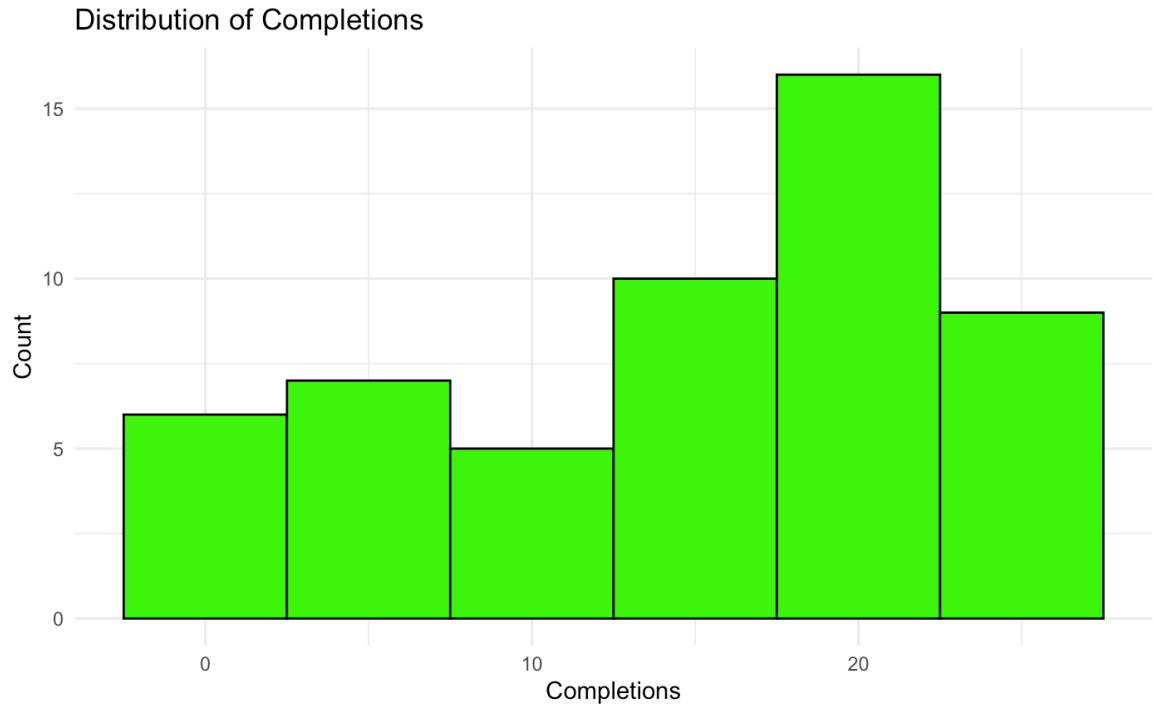
Our second dataset is a CSV file called "nflverse_df." This dataset contains more comprehensive information regarding player performance and statistical data across various games and seasons. We accessed this dataset from R package and contains useful columns such as:

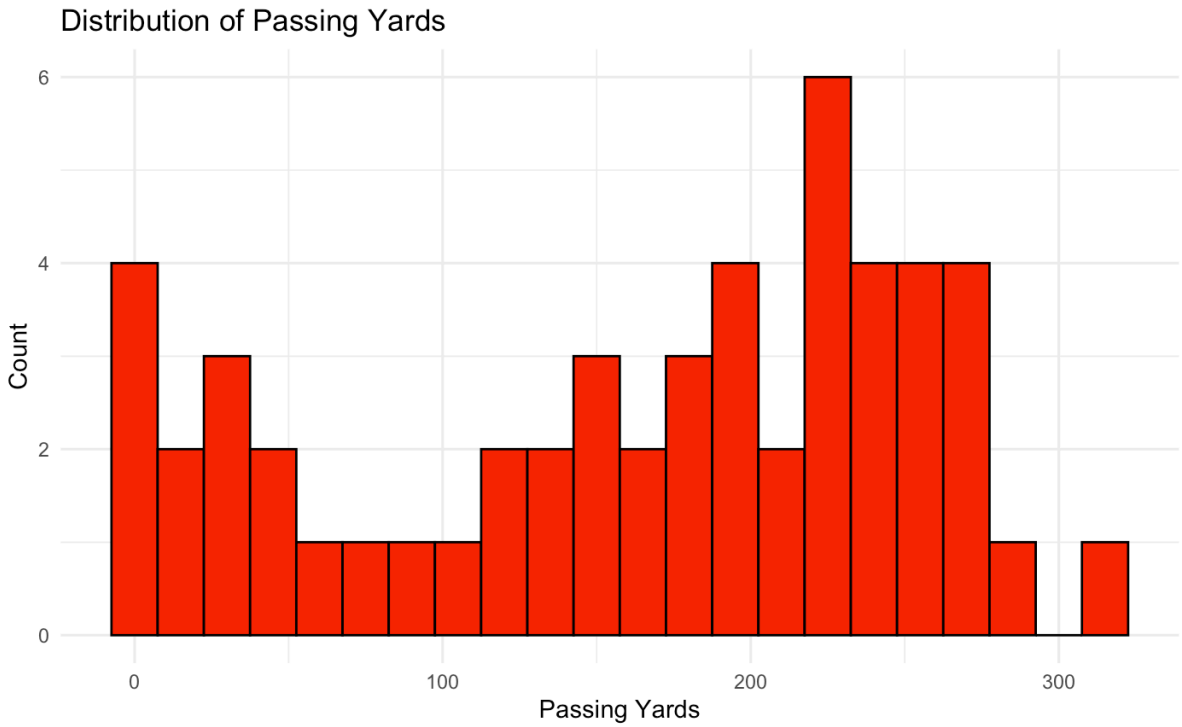
player_display_name: The full display name of the player.
position: The position the player occupies in the team.
completions: The number of completed passes.
passing_yards: Yards gained on pass plays.
rushing_yards: Yards gained when rushing with the ball

Our pre-processing step is merging the two dataset by player's name to have information regarding performance metric and salary for all the players in the dataset. In addition, the "nflverse_df" contains some player's performance of more than more seasons. For those players, we average their performance metric for each season to make their metric more clear. Last, we group players into different groups based on their position, because different positions would likely have different performance metrics, for instance, when we look at performance metrics for Running Back, we will look at rushing yards instead of passing yards.

Next, we performed EDAs for our dataset. Since there are 47 columns in the dataset and we are not using all of them as we only care about performance metrics, we have three EDAs below for “passing_yards,” “rushing_yards,” and “completions.” When performing EDA for each performance metric such as rushing yards, we did filter down the dataset to the corresponding offense and defense position.







Based on the plots above, we can see that none of them are completely normally distributed where the all three plots seem to be left-skewed to some extent. However, since the skewness is not very severe, we can use these dataset without any data transformation.

Methods

Once our data is ready, we want to find out what statistics correspond to a higher base salary, and we want to subsequently be able to predict an optimal salary for players based on their statistics that they have put up over the course of the season. Intuitively, the better the player performs, the higher their salary should be, and we would like to fit models based off of that claim for our analysis.

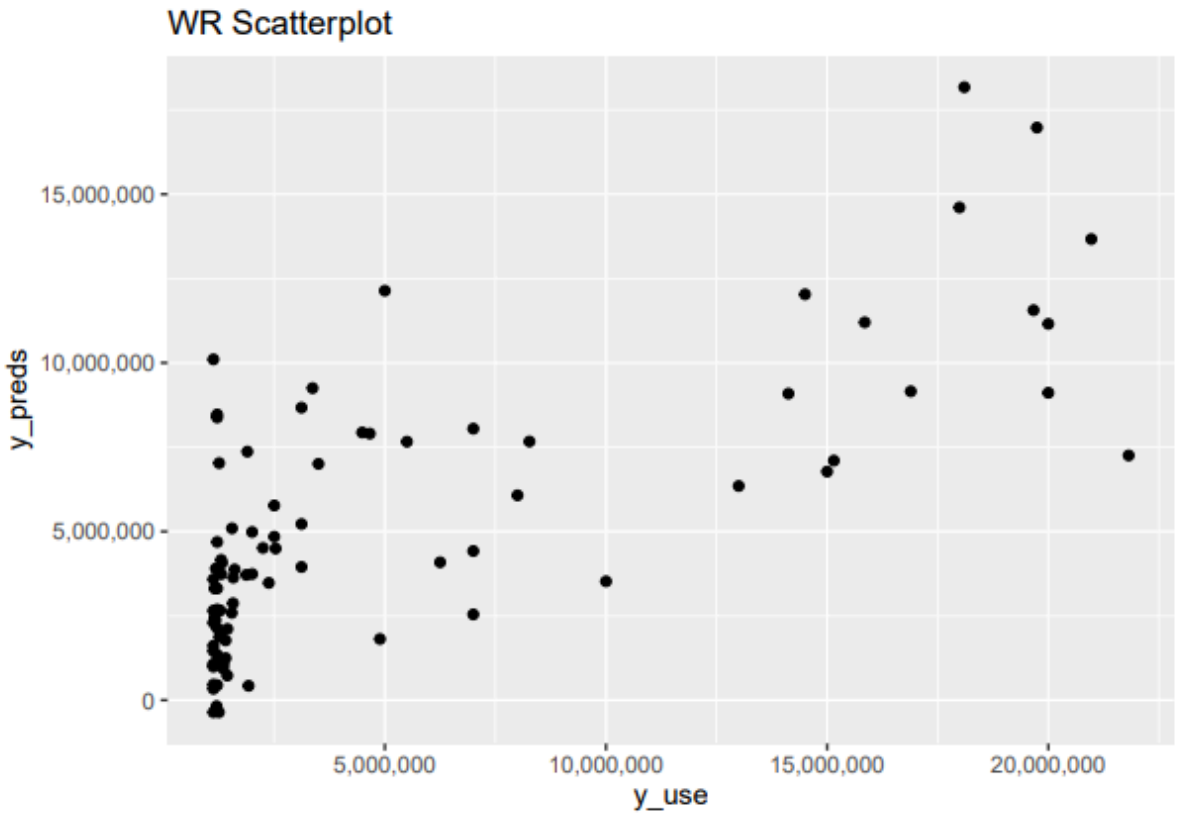
Our analysis has 2 steps, first we need to figure out what is correlated to a higher salary number, and secondly, we use our knowledge of the statistics to make projected salaries, and from these predictions, we can gain a sense of who is overpaid and who is underpaid.

Our first step involves learning about the data and eventually finding out which variables are important for determining salary. To do this, we first start with the assumption that the same statistics will not be equally valuable for each position, (for e.g: Quarterbacks would not typically have receptions [Rare Exception of Taysom Hill who receives regularly]). With this assumption in mind, we separate our data into different positions, and for each position, we use cross-validation along with Lasso regression through the cv.glmnet library in R.

Lasso helps us with regression analysis beyond a simple regression since we gain insights about which variables are the most significant, and consequently, LASSO regression works to take advantage of taking the coefficients of variables that are not relevant to the analysis towards 0. We can then use the lasso regression model in place of a normal regression model which would include statistics like receiving yards for a quarterback.

We do this for 4 position groups, QB, WR, TE, HB. These positions have the benefit of having a lot of players in these roles available for analysis. We fit LASSO models for each of these datasets, and these models give us coefficient values for important/relevant variables as well as allow us to make predictions for theoretical salary.

A scatterplot for predicted salary vs actual salary is shown below for the WR position. We have similar scatterplots for the other positions available as well.



Once we have these theoretical predictions for salary, we can make an estimation of how much a player is overpaid or underpaid. To do so, the metric we are using is % over/underpay. We calculate this by first finding the difference between the theoretical salary and the base salary, then we divide the absolute difference by the theoretical salary. By doing so, we are able to calculate a percentage metric of how much % a player is being paid over or under their deserved value.

With these, we are able to find out who our most overpaid/underpaid players are and these results are presented in the Results section.

To make a judgement of the level of uncertainty our models have, we shall attempt to quantify the standard error of the predictions that our model has given us. To do this, we shall use the statistic known as the Standard Error of the Mean or (SEM) which is calculated by dividing the standard deviation of the predictions by the square root of the number of predictions. We use this, since it gives us a measure of the average uncertainty of our 4 models, since we expect different positions to have different salary ranges and conversely, different levels of uncertainty in the model predictions. Therefore, to quantify the different uncertainties across the different predictions of the models, we use SEM.

A table with all of the SEM values for the different positions is presented below.

```
##   Position              SEM
## 1   SEM WR 422971.691023482
## 2   SEM HB 84041.5602433971
## 3   SEM TE 179205.957754454
## 4   SEM QB 559827.232093389
```

Results

Summary of model:

Our Lasso model allows us to view the most important variables that it has selected in the modeling process, and we find those by looking for the non-zero coefficients in the modeling approach.

We present the non-zero coefficients for the HB model below. Similarly we also gain access to the important variables for WR, TE, and QB as well.

```
##        (Intercept) rushing_first_downs         rushing_epa          receptions
##          1936672.99           347877.40           260127.48            34677.63
##                wopr
##          1048404.68
```

With these coefficients, we are able to see that the hb's can be evaluated strongly with rushing and receiving statistics, and seemingly to a high degree with the rushing_first_downs, and rushing_epa statistics. (rushing_first_downs = the number of first downs a hb gets while rushing, and rushing_epa = expected points added while rushing, wopr = weighted opportunity rating)

Interpretation of the model:

With our model, we generated the 10 most underpaid/overpaid players for each of the four positions by estimating their fair salaries using the Lasso Regression we trained earlier and comparing with the players actual salaries. The figure below have a graph covering results from all 4 positions:

	rank	Overpaid_HB	Underpaid_HB
1	1	Derrick Henry	David Montgomery
2	2	James Cook	Andrew Beck
3	3	Zack Moss	Michael Carter
4	4	D'Onta Foreman	Jamaal Williams
5	5	Antonio Gibson	Patrick Ricard
6	6	Brian Robinson	Miles Sanders
7	7	Rachaad White	Christian McCaffrey
8	8	Ty Johnson	Jonathan Taylor
9	9	Craig Reynolds	Alvin Kamara
10	10	Tony Pollard	Nick Chubb

17

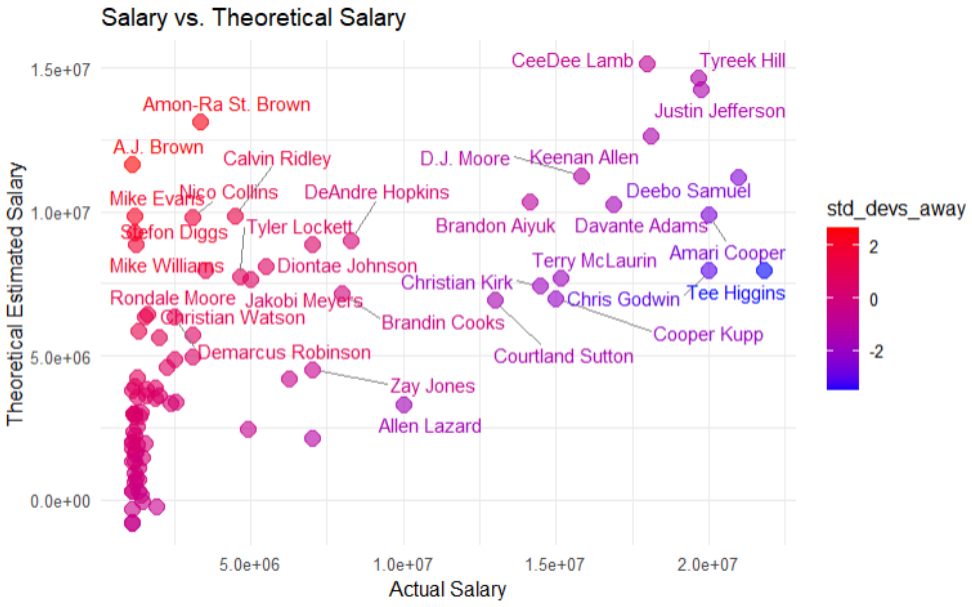
	rank	Overpaid_QB	Underpaid_QB
1	1	Joshua Dobbs	Deshaun Watson
2	2	Baker Mayfield	Geno Smith
3	3	Jalen Hurts	Matthew Stafford
4	4	Carson Wentz	Lamar Jackson
5	5	Russell Wilson	Dak Prescott
6	6	Derek Carr	Jared Goff
7	7	Tim Boyle	Tua Tagovailoa
8	8	Jimmy Garoppolo	Taysom Hill
9	9	Taylor Heinicke	Daniel Jones
10	10	Mitchell Trubisky	Kyler Murray

	rank	Overpaid_TE	Underpaid_TE
1	1	George Kittle	Tyler Conklin
2	2	Sam LaPorta	Drew Sample
3	3	Trey McBride	Darren Waller
4	4	Dallas Goedert	T.J. Hockenson
5	5	Luke Musgrave	Juwan Johnson
6	6	Hunter Henry	Foster Moreau
7	7	Jonnu Smith	Travis Kelce
8	8	Tanner Hudson	Cole Kmet
9	9	Dalton Kincaid	Mo Alie-Cox
10	10	Zach Ertz	Josh Oliver

	rank	Overpaid_WR	Underpaid_WR
1	1	A.J. Brown	Cooper Kupp
2	2	Mike Evans	Amari Cooper
3	3	Stefon Diggs	Juwann Winfree
4	4	Mike Williams	JuJu Smith-Schuster
5	5	Curtis Samuel	Allen Lazard
6	6	Kendrick Bourne	Tee Higgins
7	7	K.J. Osborn	Zach Pascal
8	8	George Pickens	Velus Jones
9	9	Jameson Williams	Gunner Olszewski
10	10	Greg Dortch	Devin Duvernay

Uncertainty Estimate:

Here, we are using the absolute difference between the players' actual salary and the theoretical salary we estimated using the model to do the estimation. The way we account for uncertainty in our estimation is by measuring the standard deviation of such difference as shown in the graph below for the wide receiver position, with larger the absolute value of standard deviation signaling the greater confidence that we have for a mismatched salary.



From our SEM table above, we also notice that our QB position has the highest SEM and the HB position has the lowest SEM. This can be explained by the fact that in the vast majority of cases, the QB will be paid considerably more than the HB, and the SEM for WR and TE are along expectations where the SEM grows as the average pay of the position grows as well.

Discussion

Based on the result shown above, we showed that our model is able to utilize the *nflverse* dataset to analyze players salaries. However, we do recognize that there are several players on our overpaid/underpaid list that seem unreasonable. We attribute the reason for the discrepancy to the limitation in our dataset, which only contains performance based data.

In the future, we aim to follow up this study by incorporating data related with player's physical conditions (such as age, height, weight, injuries) and their business worth (such as social media popularity) to provide a more accurate evaluation of their worth.

Code Appendix

https://github.com/taksh-gosw/36-660-Project