Midterm Data Project

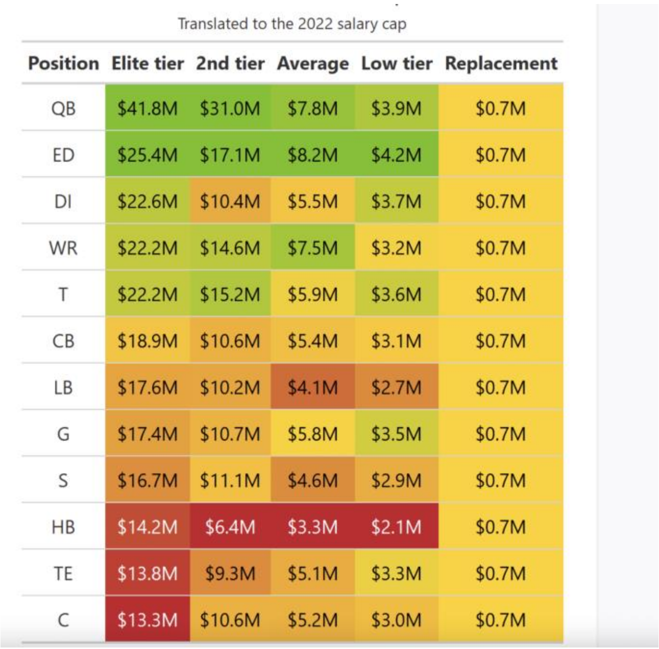
My data was collected from Pro Football Focus (PFF), the premier advanced stats source for the NFL. I used data from the 2022-2023 NFL season to look at different positions' advanced stats, quarterback (QB) and running back (RB), and their correlations to whether or not their team made the playoffs. The following stats were used for QB[1]: accuracy percentage, average depth of target, average time to throw, turnover worthy play rate, and passing yards per attempt. For RB[2], the following stats were used: breakaway percentage, yards after contact per rush attempt, and yards per rushing attempt. These statistics were compiled in a downloadable CSV file and imported into Python.

In the NFL, positional value is a widely controversial topic. Some teams refuse to draft running backs high. For example, the Eagles have not drafted a running back in the first round in over 35 years[3]. On the other hand, both the Falcons and Lions drafted a running back in the first round in the 2023 draft.[4] This topic is particularly controversial given the impact of running backs on team success, but also given their general popularity with fans. Outside of quarterbacks, more jerseys are bought by fans for running backs than any other position. In addition, running backs are the breadwinners in fantasy football, a huge outlet for fan engagement.

In 2020, PFF designed a model to rank positional values.[5] In their rankings, they found that running back was the 2nd least important position in football based on "wins above replacement." Essentially, replacing a top player at running back with an average one does not hurt you much. Unsurprisingly, quarterback was the most important position. The average "elite" quarterback in the NFL made $41.8M last year, the most in the NFL per position. To contrast, the average "elite" running back made $14.2M, the 3rd *least* of any position. Running backs also made the least money of any position with players in the "average" tier.[6]

[1] https://premium.pff.com/nfl/positions/2022/REGPO/passing?position=QB,NQB
[2] https://premium.pff.com/nfl/positions/2022/REGPO/rushing
[3] https://www.profootballnetwork.com/list-of-philadelphia-eagles-first-round-nfl-draft-picks/
[4] https://www.espn.com/nfl/draft/rounds
[5] https://www.pff.com/news/nfl-using-pro-adjusted-wins-above-average-to-examine-positional-value-in-the-nfl-draft

[6] https://www.pff.com/news/draft-surplus-value-of-each-position-in-the-nfl-draft



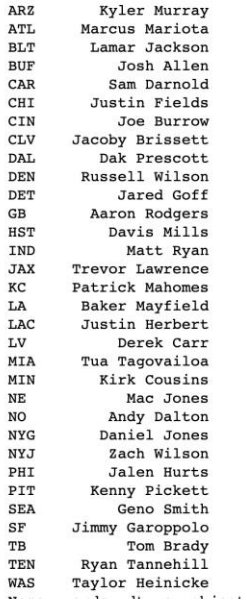
Translated to the 2022 salary cap

| Position | Elite tier | 2nd tier | Average | Low tier | Replacement |
|---|---|---|---|---|---|
| QB | $41.8M | $31.0M | $7.8M | $3.9M | $0.7M |
| ED | $25.4M | $17.1M | $8.2M | $4.2M | $0.7M |
| DI | $22.6M | $10.4M | $5.5M | $3.7M | $0.7M |
| WR | $22.2M | $14.6M | $7.5M | $3.2M | $0.7M |
| T | $22.2M | $15.2M | $5.9M | $3.6M | $0.7M |
| CB | $18.9M | $10.6M | $5.4M | $3.1M | $0.7M |
| LB | $17.6M | $10.2M | $4.1M | $2.7M | $0.7M |
| G | $17.4M | $10.7M | $5.8M | $3.5M | $0.7M |
| S | $16.7M | $11.1M | $4.6M | $2.9M | $0.7M |
| HB | $14.2M | $6.4M | $3.3M | $2.1M | $0.7M |
| TE | $13.8M | $9.3M | $5.1M | $3.3M | $0.7M |
| C | $13.3M | $10.6M | $5.2M | $3.0M | $0.7M |

So there seems to be a dilemma here. Data analysts generally agree that the running back position is unimportant compared to other positions. However, some teams are still using high-leverage draft picks to secure them. NFL teams must balance fan engagement with strategy. They have to sell tickets but win games at the same time. My data cannot make this decision for teams. Merely, my data analysis is an attempt to corroborate what other data scientists believe—that the running back position is rather unimportant in generating success, especially as compared to the quarterback position, widely regarded as the most important position in the NFL.

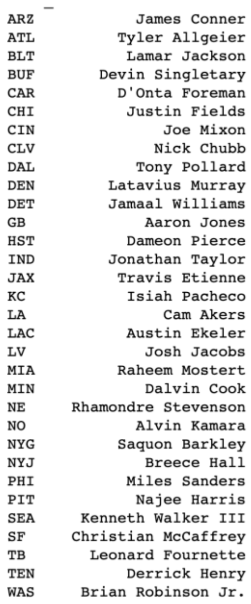
My data analysis is rooted in finding the correlations between advanced stats and whether or not a team made the playoffs. Therefore, my first step was to create a dictionary as to whether a team made the playoffs (after some thorough data cleaning, of course). After adding said dictionary to my overall dataframe, I had more cleaning to do. It is not particularly relevant to my research how backups performed, so I cleaned the dataframe to only include the most prominent player (most passing/rushing yards) from each team at their respective position:

Figure 1: Quarterbacks



```
ARZ        Kyler Murray
ATL      Marcus Mariota
BLT        Lamar Jackson
BUF          Josh Allen
CAR         Sam Darnold
CHI        Justin Fields
CIN          Joe Burrow
CLV     Jacoby Brissett
DAL         Dak Prescott
DEN      Russell Wilson
DET          Jared Goff
GB         Aaron Rodgers
HST          Davis Mills
IND            Matt Ryan
JAX      Trevor Lawrence
KC       Patrick Mahomes
LA         Baker Mayfield
LAC       Justin Herbert
LV            Derek Carr
MIA      Tua Tagovailoa
MIN         Kirk Cousins
NE             Mac Jones
NO           Andy Dalton
NYG         Daniel Jones
NYJ          Zach Wilson
PHI          Jalen Hurts
PIT        Kenny Pickett
SEA           Geno Smith
SF      Jimmy Garoppolo
TB             Tom Brady
TEN       Ryan Tannehill
WAS      Taylor Heinicke
```

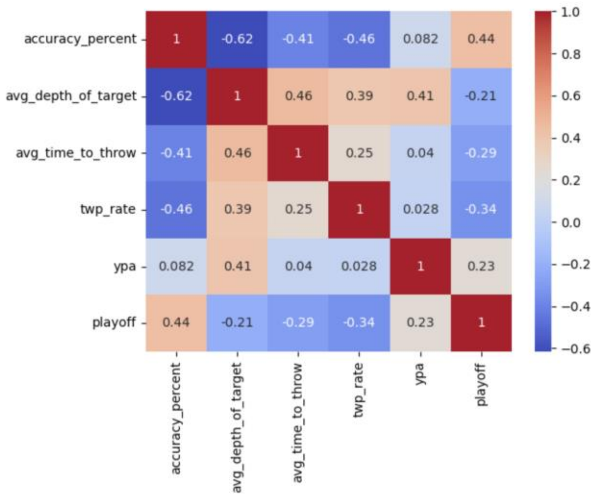
Figure 2: Running Backs



```
ARZ              James Conner
ATL            Tyler Allgeier
BLT             Lamar Jackson
BUF          Devin Singletary
CAR            D'Onta Foreman
CHI             Justin Fields
CIN                 Joe Mixon
CLV               Nick Chubb
DAL              Tony Pollard
DEN            Latavius Murray
DET            Jamaal Williams
GB                Aaron Jones
HST             Dameon Pierce
IND            Jonathan Taylor
JAX            Travis Etienne
KC              Isiah Pacheco
LA                  Cam Akers
LAC             Austin Ekeler
LV                Josh Jacobs
MIA            Raheem Mostert
MIN               Dalvin Cook
NE       Rhamondre Stevenson
NO               Alvin Kamara
NYG            Saquon Barkley
NYJ               Breece Hall
PHI             Miles Sanders
PIT               Najee Harris
SEA        Kenneth Walker III
SF        Christian McCaffrey
TB          Leonard Fournette
TEN             Derrick Henry
WAS        Brian Robinson Jr.
```

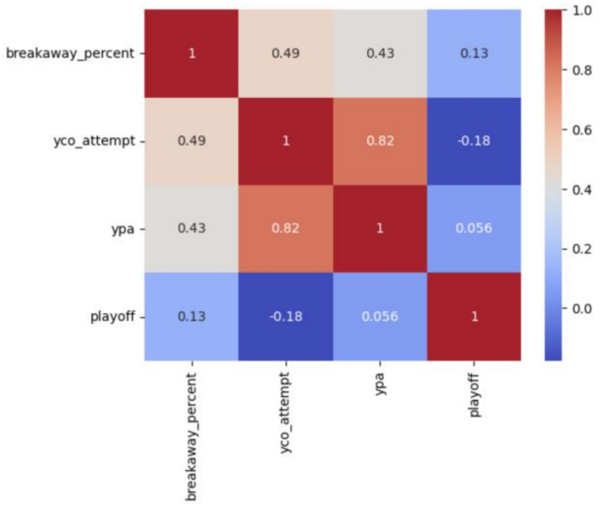
Now that I had my data clean, I was able to run the proper tests. I started with correlation heatmaps to reveal a general sense of the relationships between these variables.

Figure 3: Quarterback Advanced Stats vs. "Playoff"



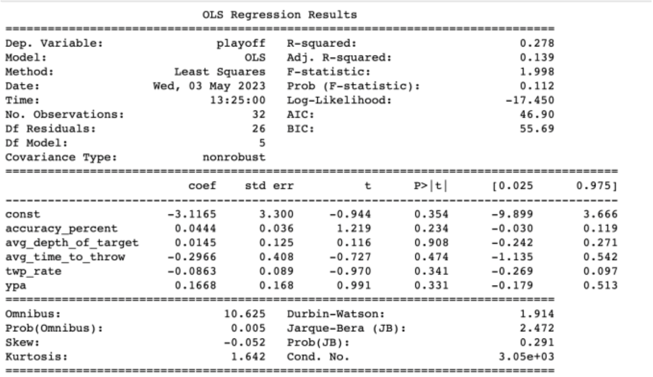
The data shows that the strongest quarterback correlations to whether a team makes the playoffs, "playoff", are accuracy percent and turnover worthy plays rate (TWP). Essentially, teams need their quarterback to complete passes at a high rate and not turn the ball over. At first glance this feels obvious, but it is more useful as a source of comparison. As our heatmap shows, these stats are more important than depth of target and yards per attempt. Completing long passes helps, but ball security and accuracy are more important.

Figure 4: Running Back Advanced Stats vs. “Playoff”



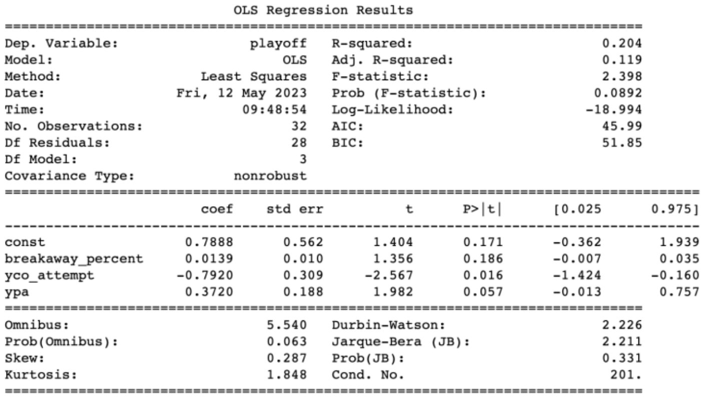
Here, the correlation coefficients are much smaller. In fact, the highest correlation to playoff, yards after contact per rushing attempt, is smaller than the lowest correlation coefficient for the quarterback stats above. Already, the data shows that running back value fails in comparison to quarterback value. Interestingly, running back rushing yards per attempt brings almost no value with a coefficient of 0.056.

Next, I ran logistic regression to see how well these statistics predicted whether a team made the playoffs.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                playoff   R-squared:                       0.278
Model:                            OLS   Adj. R-squared:                  0.139
Method:                 Least Squares   F-statistic:                     1.998
Date:                Wed, 03 May 2023   Prob (F-statistic):              0.112
Time:                        13:25:00   Log-Likelihood:                -17.450
No. Observations:                  32   AIC:                             46.90
Df Residuals:                      26   BIC:                             55.69
Df Model:                           5
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const                 -3.1165      3.300     -0.944      0.354      -9.899       3.666
accuracy_percent       0.0444      0.036      1.219      0.234      -0.030       0.119
avg_depth_of_target    0.0145      0.125      0.116      0.908      -0.242       0.271
avg_time_to_throw     -0.2966      0.408     -0.727      0.474      -1.135       0.542
twp_rate              -0.0863      0.089     -0.970      0.341      -0.269       0.097
ypa                    0.1668      0.168      0.991      0.331      -0.179       0.513
==============================================================================
Omnibus:                       10.625   Durbin-Watson:                   1.914
Prob(Omnibus):                  0.005   Jarque-Bera (JB):                2.472
Skew:                          -0.052   Prob(JB):                        0.291
Kurtosis:                       1.642   Cond. No.                     3.05e+03
==============================================================================
```

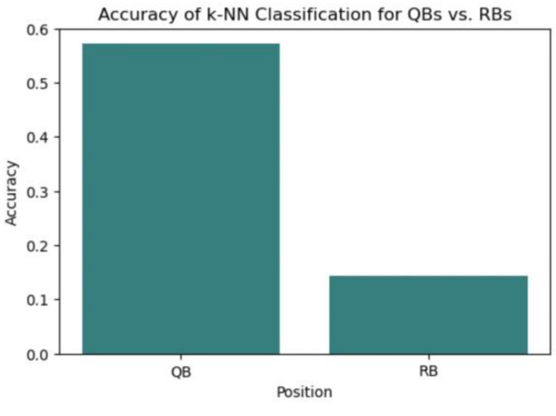
In the context of football, it's important to note that no single statistical measure can fully capture the performance of a team, as it involves a complex interplay of multiple factors. However, in the regression analysis of quarterback performance, the R-squared value of 0.278 indicates that 27.8% of the variation in the dependent variable (playoff success) can be explained by the independent variables (accuracy, depth of target, time to throw, twp rate, and yards per attempt) included in the model. Although none of the coefficients are statistically significant at conventional levels, the relatively high R-squared value suggests that these variables still matter to a certain extent, especially when compared to other positions on the team.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                playoff   R-squared:                       0.204
Model:                            OLS   Adj. R-squared:                  0.119
Method:                 Least Squares   F-statistic:                     2.398
Date:                Fri, 12 May 2023   Prob (F-statistic):             0.0892
Time:                        09:48:54   Log-Likelihood:                -18.994
No. Observations:                  32   AIC:                             45.99
Df Residuals:                      28   BIC:                             51.85
Df Model:                           3
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 0.7888      0.562      1.404      0.171      -0.362       1.939
breakaway_percent     0.0139      0.010      1.356      0.186      -0.007       0.035
yco_attempt          -0.7920      0.309     -2.567      0.016      -1.424      -0.160
ypa                   0.3720      0.188      1.982      0.057      -0.013       0.757
==============================================================================
Omnibus:                        5.540   Durbin-Watson:                   2.226
Prob(Omnibus):                  0.063   Jarque-Bera (JB):                2.211
Skew:                           0.287   Prob(JB):                        0.331
Kurtosis:                       1.848   Cond. No.                         201.
==============================================================================
```

This OLS regression analysis shows that the model with three independent variables (breakaway percent, yards after contact per attempt, and yards per attempt) explains 20.4% of the variance in playoff performance. While none of the independent variables are statistically significant at the 0.05 level, the coefficients for yards after contact per attempt and yards per attempt are both positive, indicating that these variables have a positive relationship with playoff performance. The coefficient for breakaway percent is positive but not statistically significant. Therefore, while the model does not have strong explanatory power, the included variables are still important predictors for playoff performance.

Finally, I used K-Nearest Neighbors to compare QB and RB accuracy at predicting "playoff"



K-Nearest Neighbors (KNN) is a simple machine learning algorithm that is commonly used for classification problems. It works by identifying the k nearest data points to a new data point and classifying the new data point based on the class that is most common among its k nearest neighbors. In this case, I used KNN to classify whether or not a team made the playoffs based on the performance of either their quarterback or running back. For the RB data, I selected four important stats, including breakaway percent, yards after contact per rushing attempt, yards per attempt, and playoff status, and trained a KNN classifier on this data. The classifier achieved an accuracy of 14.29%, indicating that it was not very effective at predicting playoff success based on these RB stats. For the QB data, I selected five important stats, including accuracy percent, average depth of target, average time to throw, turnover worthy plays rate, yards per attempt, and playoff status. The KNN classifier trained on this data achieved an accuracy of 57.14%, indicating that it was much more effective at predicting playoff success based on these QB stats. Overall, this suggests that QB performance is a more important factor than RB performance in determining playoff success in the NFL. While RB stats may contribute to a team's overall performance, they do not appear to be as strongly correlated with making the playoffs as QB stats.

I've observed a dilemma in the NFL with respect to the significance of the running back position relative to other positions. While some teams still use high draft picks to secure running backs, most data experts agree that the position is relatively unimportant in generating success, particularly when compared to the quarterback position, which is widely considered the most critical position in the NFL. To support this assertion, I've conducted data analysis that shows the strongest correlations between QB and RB performance and whether a team makes the playoffs. The data indicates that completing passes at a high rate and avoiding turnovers are the most critical factors in determining playoff success. Completing long passes helps, but ball security and accuracy are more important. On the other hand, the data shows that the value of RBs fails in comparison to that of QBs. Running back rushing yards per attempt brings almost no value, and the highest correlation to playoff success, yards after contact per rushing attempt, is smaller than the lowest correlation coefficient for the quarterback stats. Using the K-Nearest Neighbors (KNN) machine learning algorithm to classify whether a team made the playoffs based on the performance of either their quarterback or running back, the KNN classifier trained on quarterback data achieved an accuracy of 57.14%, indicating that it was much more effective at predicting playoff success based on these QB stats as compared to the 14.29% for RB. In conclusion, while running back stats may contribute to a team's overall performance, they do not appear to be as strongly correlated with making the playoffs as quarterback stats. NFL teams must balance fan engagement with strategy and win games at the same time, and while my data analysis cannot make this decision for them, it provides evidence that they should prioritize investing in their quarterback position over their running back position.