Project Title:
Drafting for Dollars: Isolating the ROI of NFL Draft Picks on Team Performance
Team Members: Ashton Mehta, Bo Han, Govind Rachapudi, Viren Halaharivi

Research Question:
How well can we predict a player's value based on their pre-draft performance metrics?

Motivation:
We chose to focus on building a draft model to determine ROI of draft picks based on pre-draft performance. We will quantify this pre-draft performance by using player scouting grades and weighted career approximate value of players drafted. What makes our research question interesting is that the primary objective of NFL teams during the draft is to draft players that can positively impact the team in the present and years into the future. Player value can depend on position needs or just plain generational talent. However, finding a reliable method to predict present and future impact by players or ROI on draft picks is required to make informative decisions that can impact team success for years to come. This method can utilize pre-draft performance to predict this ROI.

Data Sources:

1. **Draft Information:**
   We used first- and second-round NFL Draft data from Pro-Football-Reference.com for our player and team selections. This dataset provided detailed historical draft results that served as the foundation for our project.
2. **Combine Data:**
   Combine performance metrics (40-yard dash, bench press, vertical jump, broad jump, 3 Cone, Shuttle) were also retrieved from Pro-Football-Reference.com. These physical measurements were included to help evaluate athletic potential and project future player performance.
3. **Pre-Draft Prospect Grading:**
   We used prospect grades from NFL.com for players drafted between 2014 and 2024. These grades range from 5.6 ("Practice Squad Player") to 8.0 ("Perfect Prospect") and provided standardized evaluations of players' perceived potential entering the NFL.
4. **Career College Statistics:**
   Career college statistics for prospects (such as rushing yards, passing yards, tackles, etc.) were parsed from Pro-Football-Reference.com. These statistics were incorporated into our model to better predict NFL success. We retrieved relevant statistics for the relevant positions.
5. **Post-Draft Performance Metrics:**
   We used **Approximate Value (AV)** to assess player success after being drafted. AV was the primary outcome variable that we aimed to predict with our model.
6. **Project Goal:**
   Using the combination of pre-draft evaluations (combine data, prospect grades, and

college statistics) and historical draft trends, our objective was to build a predictive model to forecast a player's **AV** in the NFL based on their profile at the time of the draft.

Data Collection and Preparation:

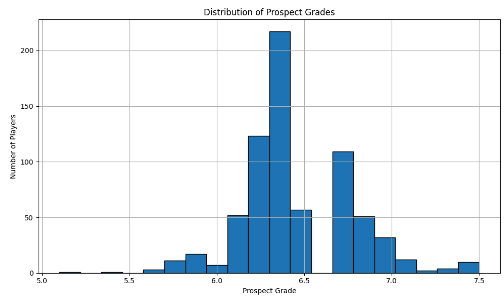
The goal for the draft model was to be holistic. Existing approaches have utilized multiple data sources for calculating a “draft score”, which ideally correlates to career success in the NFL. Taking the AWS Next Gen Stats model, for instance, the data sources include various college statistics and combined data. Once aggregated, the data is both standardized and transformed through a series of models to produce the corresponding output. Our model took inspiration not only from the use of statistics across various data sources but also in separating the models by training according to position. The reason behind this decision lies in the observation that different positions' performance are measured using different statistics. For instance, the passer rating for the quarterback cannot be effectively used to determine the performance for running backs. In summary, the data used for our model included NFL combine data such as height, weight, and bench, college performance data separated by position, and NFL prospect grades.

We used the publicly available data from NFL.com to aggregate prospect grades. For each player drafted between 2014 and 2024, the timespan used as a constraint in our modeling data, a number ranging from 5.6 to 8.0 was given. The number on the lower end indicates a candidate who would, as described by scouts, likely be on a practice squad. However, the number on the higher end of the scale would indicate a “perfect candidate” for the NFL. One observation that could be made was that over time, the grades became more and more precise. Each grade was manually queried and entered in our database for further analysis.

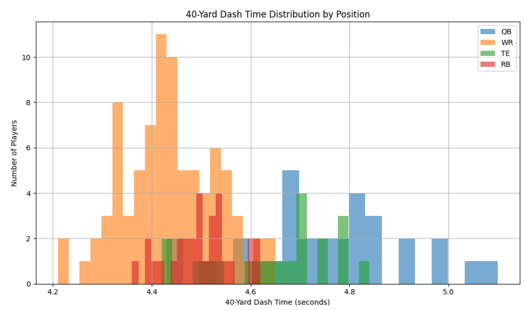
To collect career statistics for each NFL prospect in our database, separated by position, we used a Selenium-based web scraping approach. In our research, we were unable to find a data source that was easy to interact with and aggregate data in bulk. There were no free existing APIs or packages that were available to use. However, we were able to locate a website, sports-reference.com, that provided individual college career statistics for NFL prospects. Given the standardized URLs and structure in the HTML code for each page, we were able to pivot to using Selenium, a python library which allows developers to interact with web pages programmatically. Using a custom script, a URL corresponding to each player was scraped and the corresponding data was entered into the database. The scope of our model limited statistics to career totals according to position, rather than per-year data.

Finally, the NFL combine data was procured through pro-football-reference.com. Statistics such as height, weight, bench, 40-yard dash, and vertical jump were selected for each player. The process of scraping the data was straightforward, as the numbers were available in bulk tables. To add them to our database, we simply copied and pasted the numbers going back to 2014 in a separate spreadsheet. The three data sources were then joined together by the player name and the “target” Approximate Value (AV) variable was added for each player. The final dataset, separated by position, thus included college total statistics, prospect grades, NFL combine statistics, and career AV.

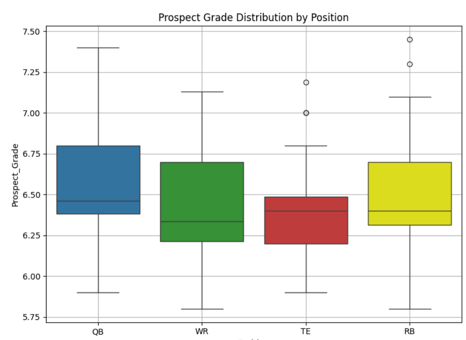
Methodology: Exploratory Data Analysis and Visualization



- Distribution appears to have a normal shape
- Unimodal with a peak grade of 6.4



- WRs dominate the fastest 40 yard times and cluster around 4.3-4.5 seconds
- QBs show wider and slower distributions typically above 4.7 seconds
- RBs and TEs fall in between, with RBs skewing faster than TEs



- QBs have the highest median and widest spread in prospect grades, indicating varied evaluations
- TEs have the lowest median and tighter distribution, with several outliers above 7.0
- WRs and RBs have similar medians around 6.4

Results:

Data Modeling and ROI Calculation:

- Created an ROI prediction model based on drAV (approximate value) and draft round.
- Trained a linear regression model with testing and training sets to measure the Mean Squared Error and R-Squared of ROI values for quarterbacks to further evaluate the accuracy of the model.

```python
predictors_roi = [
    'DrAV_nfl', 'Rnd_nfl', 'Pass_Cmp_nfl', 'Pass_Att_nfl', 'Pass_Yds_nfl',
    'Pass_TD_nfl', 'Pass_Int_nfl',
    'G_college', 'Cmp_college', 'Att_college', 'Yds_college', 'TD_college'
]
target_roi = 'ROI'

# Calculate ROI
qbs_merged['ROI'] = qbs_merged['DrAV_nfl'] / qbs_merged['Rnd_nfl']

# Handle missing values
qbs_merged_filled = qbs_merged[predictors_roi + [target_roi]].fillna(0)

# Check if data is empty
if qbs_merged_filled.empty:
    raise ValueError('The dataset is empty after handling missing values.')

# Split into X and y
X_roi = qbs_merged_filled[predictors_roi]
y_roi = qbs_merged_filled[target_roi]

# Train-test split
X_train_roi, X_test_roi, y_train_roi, y_test_roi = train_test_split(X_roi, y_roi, test_size=0.2, random_state=42)

# Model training
roi_model = LinearRegression()
roi_model.fit(X_train_roi, y_train_roi)

# Predictions and evaluation
y_pred_roi = roi_model.predict(X_test_roi)
mse_roi = mean_squared_error(y_test_roi, y_pred_roi)
r2_roi = r2_score(y_test_roi, y_pred_roi)

print(f'Mean Squared Error (ROI): {mse_roi:.2f}')
print(f'R-squared (ROI): {r2_roi:.2f}')
```

```
Mean Squared Error (ROI): 31.33
R-squared (ROI): 0.97
```

- Essentially, ROI was calculated by approximate value (drAV) by NFL draft round for quarterbacks.
- The R^2 is high but the dataset is small which means that the model can be overfitting on the data due to it being a linear model.
- To fix this we can increase the amount of rows in the dataset by adding more players to the dataset.

```python
# Example: Run ROI prediction for an actual QB
example_qb = qbs_merged.iloc[1]  # Select the first QB as an example
example_data = example_qb[predictors_roi].values.reshape(1, -1)
predicted_roi = roi_model.predict(example_data)[0]
print(f"Player: {example_qb['Player_college']}")
print(f"Predicted ROI: {predicted_roi:.2f}")
print(f"Actual ROI: {example_qb['ROI']:.2f}")
```

```
Player: Johnny Manziel
Predicted ROI: 4.56
Actual ROI: 4.00
```

- Using my model I predicted the ROI of Johnny Manziel and it seems that my model is relatively accurate at predicting the Return on Investment of a Quarterback.

```python
# Example: Run ROI prediction for an actual QB
example_qb = qbs_merged.iloc[5]
example_data = example_qb[predictors_roi].values.reshape(1, -1)
predicted_roi = roi_model.predict(example_data)[0]
print(f"Player: {example_qb['Player_college']}")
print(f"Predicted ROI: {predicted_roi:.2f}")
print(f"Actual ROI: {example_qb['ROI']:.2f}")
```

```
Player: Jameis Winston
Predicted ROI: 56.12
Actual ROI: 54.00
```

- This is another prediction of ROI for the QB Jameis Winston and it is relatively accurate.

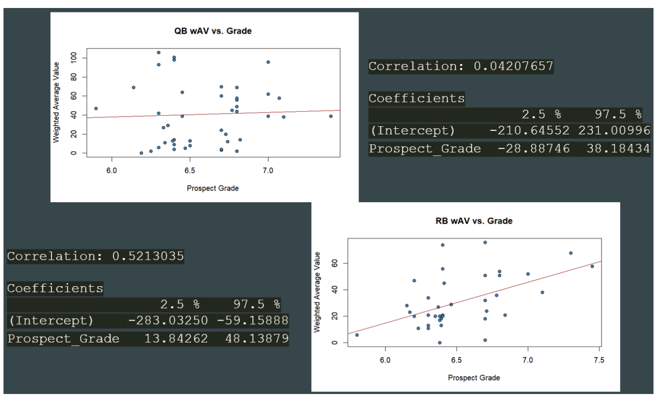
```python
# Example: Run ROI prediction for an actual QB
example_qb = qbs_merged.iloc[3]
example_data = example_qb[predictors_roi].values.reshape(1, -1)
predicted_roi = roi_model.predict(example_data)[0]
print(f"Player: {example_qb['Player_college']}")
print(f"Predicted ROI: {predicted_roi:.2f}")
print(f"Actual ROI: {example_qb['ROI']:.2f}")
```

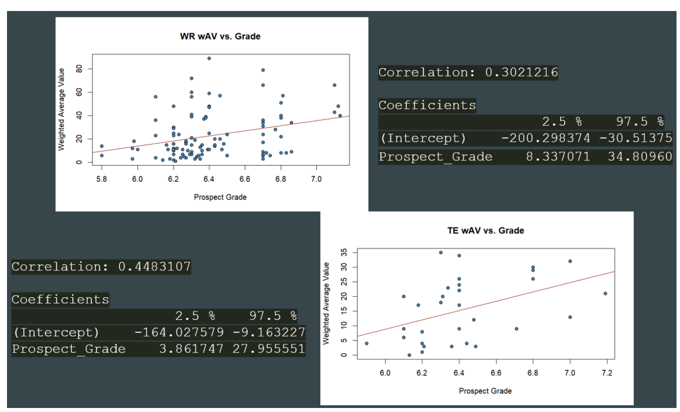
```
Player: Derek Carr
Predicted ROI: 49.11
Actual ROI: 40.00
```

- Although most of my predictions have been accurate, some of my predictions have been off most likely due to the limits of linear regression and the use of a high amount of predictors in a relatively small dataset.

Simple and Multiple Linear Regression:
We chose to analyze drafting strategies using four methods: using NFL prospect grades, the eye test, NFL combine results, and college performance.

1. NFL Prospect Grades:
   We used simple linear regression to model NFL prospect grades on weighted average value (wAV) that we received from Pro-Football-Reference.com. We repeated this model on data for quarterbacks, running backs, wide receivers, and tight ends. Running backs, wide receivers, and tight ends had moderately positive correlations between wAV and prospect grade, however for quarterbacks the correlation was almost zero. This is also supported by the coefficients of the model as well as the regression line. The 95% confidence intervals for the model coefficients all have significance except for quarterbacks, so using prospect grades may be a viable strategy to evaluate player value and draft offensive skill positions other than quarterbacks.





2. The Eye Test:
   The eye test is analyzing prospects using their physical traits, or if a prospect “looks” like an NFL player that will bring value to the field. We used multiple linear regression to predict wAV, regressing on height and weight. We again repeated this model for the four

offensive skill position data described previously. Similar to the prospect grades, there was almost no correlation between height and weight with wAV for quarterbacks. For running backs there was a stronger, low-to-moderate positive correlation. Wide receivers had low positive correlations, and for tight ends, there surprisingly were moderately negative correlations. These negative correlations may be due to the fact that tight ends who are too tall and too heavy may be ineffective as receivers and more effective as "linemen". The 95% confidence intervals for the model coefficients showed almost no significance, which may point towards the fact that using the eye test as a strategy to evaluate player value and draft is unreliable.

QBs Correlation

	wAV
wAV	1.00000000
Ht_inch	-0.07053628
Wt	0.01388233

QBs Coefficients

	2.5 %	97.5 %
(Intercept)	-253.8068077	578.297081
Ht_inch	-9.9534819	4.748394
Wt	-0.8818825	1.549615

RBs Correlation

	wAV
wAV	1.0000000
Ht_inch	0.3657417
Wt	0.2576134

RBs Coefficients

	2.5 %	97.5 %
(Intercept)	-585.3635716	28.1910371
Ht_inch	-1.2529985	9.5491664
Wt	-0.6687084	0.8108294

WRs Correlation

	wAV
wAV	1.0000000
Ht_inch	0.1294108
Wt	0.2404916

WRs Coefficients

	2.5 %	97.5 %
(Intercept)	-134.73269863	156.2722415
Ht_inch	-3.44495922	1.7195364
Wt	0.02939661	0.7142921

TEs Correlation

	wAV
wAV	1.0000000
Ht_inch	-0.2512285
Wt	-0.4307296

TEs Coefficients

	2.5 %	97.5 %
(Intercept)	-297.734832	423.831293719
Ht_inch	-4.571364	7.162238696
Wt	-1.161648	-0.005744843

3. NFL Combine Results:

   Next, we decided to analyze how well prospects do at NFL combine drills can affect player value. We again used multiple linear regression to predict wAV based on results from the 40 yard dash, bench press, vertical jump, standing broad jump, 3 cone drill, and 20 yard shuttle for each of the four offensive skill position data. Sticking to the trend, quarterbacks had low correlations between these combine drills and wAV. The 20 yard shuttle stood out as a positive factor for running backs, but bench was also shown to be a negative factor. This may be because an increase in bench press may decrease the speed and agility that is necessary for running backs. There were low associations between combine drills and wAV for wide receivers, and all combine factors appeared to be moderately correlated with wAV for tight ends. However, the 95% confidence intervals show almost no significance, except for the 20 yard shuttle for running backs.

QBs Correlation

	wAV
wAV	1.00000000
X40yd	-0.18699488
Vertical	-0.18487976
Broad.Jump	0.13515182
X3Cone	-0.27242496
Shuttle	-0.08501118

QBs Coefficients

	2.5 %	97.5 %
(Intercept)	-1554.140760	1473.144057
X40yd	-167.050171	137.500985
Vertical	-17.676249	2.736667
Broad.Jump	-2.827353	10.043808
X3Cone	-189.461181	176.374250
Shuttle	-138.717180	149.938103

RBs Correlation

	wAV
wAV	1.00000000
X40yd	-0.31590019
Vertical	-0.12591185
Bench	-0.51297634
Broad.Jump	0.03855318
X3Cone	0.14450343
Shuttle	0.62706026

RBs Coefficients

	2.5 %	97.5 %
(Intercept)	-2155.3764424	40.8485306
X40yd	-164.2847684	280.9648508
Vertical	-6.9753457	5.6032903
Bench	-5.9336462	-0.4897852
Broad.Jump	0.2361109	6.9216519
X3Cone	-87.8197235	78.3734834
Shuttle	26.6611368	219.6177458

WRs Correlation

	wAV
wAV	1.000000000
X40yd	0.220583280
Vertical	-0.049760095
Bench	-0.004105177
Broad.Jump	-0.185324474
X3Cone	-0.017251762
Shuttle	-0.162506017

WRs Coefficients

	2.5 %	97.5 %
(Intercept)	-712.945740	440.498218
X40yd	-40.603425	180.578108
Vertical	-3.648702	5.483500
Bench	-2.255223	2.734884
Broad.Jump	-2.933233	1.065541
X3Cone	-43.368257	76.553822
Shuttle	-112.929579	27.764456

TEs Correlation

	wAV
wAV	1.0000000
X40yd	-0.7171129
Vertical	0.7030644
Bench	-0.6521035
Broad.Jump	0.7717722
X3Cone	-0.7425957
Shuttle	-0.6431500

TEs Coefficients

	2.5 %	97.5 %
(Intercept)	-321.077192	435.546015
X40yd	-86.385913	26.930588
Vertical	-1.066858	3.120621
Bench	-2.794918	0.536656
Broad.Jump	-1.083053	2.228572
X3Cone	-36.811307	53.323416
Shuttle	-84.055899	64.050629

4. College Performance:
   Lastly, we analyzed how college performance impacts player value. For quarterback data, we regressed wAV on games played, completion rate, passing yards, touchdown rate, interception rate, and passer rating. Completion and interception rate shows to be moderately correlated, however none of the college performance factors showed significance. For running backs, wide receivers, and tight ends, we regressed wAV on games played, yards per attempt, yards per reception, touchdowns, and yards per game. For running backs, yards per reception showed to have some significance, which may show how being a dual threat as a back and a receiver contributes to player value. As expected, receiving yards per game shows moderate correlations with wAV for wide receivers and tight ends, but no significance was shown by the model coefficients at the 95% level.

QBs Correlation

	wAV
wAV	1.0000000000
G_college	-0.0004450872
Cmp.	-0.3263091534
Yds	0.0449071434
TD.	0.0562611255
Int.	0.2679070892
Rate	-0.0970800367

QBS Coefficients

	2.5 %	97.5 %
(Intercept)	-81.30102832	4.045986e+02
G_college	-1.64536496	1.730138e+00
Cmp.	-9.80562614	2.448814e+00
Yds	-0.00249161	9.191223e-03
TD.	-10.12422648	3.316641e+01
Int.	-7.34197444	3.697895e+01
Rate	-3.24390711	2.899035e+00

RBs Correlation

	wAV
wAV	1.00000000
G_college	-0.02203074
Y.A	0.29808645
Y.R	0.44765363
TD	0.08304180
Y.G	0.29505384

RBs Coefficients

	2.5 %	97.5 %
(Intercept)	-137.69902150	-1.4907201
G_college	-0.56333412	1.5532692
Y.A	-4.78705122	15.6819468
Y.R	0.52318205	6.2084243
TD	-1.43384049	0.2772517
Y.G	-0.01381816	0.7761912

WRs Correlation

	wAV
wAV	1.000000000
G_college	-0.134232664
Y.R	-0.008650581
TD	0.144762112
Y.G	0.332294915

WRs Coefficients

	2.5 %	97.5 %
(Intercept)	-50.28567673	42.8728671
G_college	-0.71488646	0.5876090
Y.R	-1.52852450	2.1608491
TD	-0.75041400	0.7056841
Y.G	0.05409109	0.6914710

TEs Correlation

	wAV
wAV	1.00000000
G_college	-0.09609525
Y.R	0.28262021
TD	0.06618878
Y.G	0.34510191

TEs Coefficients

	2.5 %	97.5 %
(Intercept)	-66.31026131	37.8824446
G_college	-0.48676320	0.7408886
Y.R	-1.49619931	4.0657696
TD	-1.23910961	0.5214242
Y.G	-0.06063402	0.6714706

6. Discussion and Conclusion:

ROI Calculation Results

Using NFL combine data, college statistics, and prospect grades, we trained a linear regression model to estimate a quarterback's post-draft ROI, which is defined as the Approximate Value (DrAV) to Draft Round ratio. One of the models for QB ROI has a high R-squared value of 0.97, indicating significant predictive potential. However, because the QB dataset is tiny, this must be read with caution. A model trained on limited data is at risk of overfitting, which occurs when a model performs well on known data but performs badly when applied to unknown values. Despite these restrictions, the model identified clear connections between pre-draft performance and NFL worth, demonstrating that indicators such as combine scores and college numbers might help inform draft decisions.

Four Method Analysis Results

Although most of the results from the model lack significance, they provide a helpful pipeline when paired with correlations to analyze player value. Using prospect grades and the NFL formula of grading players to draft appears to be a viable, somewhat reliable way to draft/evaluate offensive skill position players, disregarding quarterbacks. Using the eye test does not appear to be viable, but may just provide reassurance to management as they view game tape of players. Combine results do not appear to be reliable (except for the 20 yards shuttle for running backs) as many of the drills do not correlate to on-field play, which may be why many players choose to not participate in the combine. Lastly, college performance may possibly be viable when looking at percentage statistics like completion rate and yards per reception. However, many prospects who thrive in college (Johnny Manziel and Tim Tebow) fail to transfer that success to the main stage. A last note regarding quarterback performance evaluation is that it is highly variable. Many top prospects like the college quarterbacks mentioned before fail to have the same success against professionals, whereas many low-rated prospects (Tom Brady, Brock Purdy, and possibly Shedeur Sanders now) go on to play in Super Bowls and become some of the greatest players of all time.