Milestone 2: Project Progress Report

Project Title: Does Athleticism and Draft Priority Translate to Success in the NFL?
Project Group #: 10
Authors: William Guy, Vitush Agarwal, Praveen Manimaran
Emails: v8agarwal@ucsd.edu, pmanimaran@ucsd.edu, wguy@ucsd.edu

---

Background

The NFL Combine is a crucial event for evaluating athlete performance before the draft, but does it truly predict success in the league? Our project seeks to determine if a player's athletic performance and draft position correlate with career success, using fantasy football points as a proxy for performance.

Dataset Information

- **Dataset Name:** NFL Player Statistics Dataset (2004-Present), NFL Stats 2012-2023
- **Source:** Kaggle (https://www.kaggle.com/datasets/toddsteussie/nfl-play-statistics-dataset-2004-to-present) (https://www.kaggle.com/datasets/philiphyde1/nfl-stats-1999-2022?select=yearly_player_data.csv)

Data Pipeline

Purpose: To merge and preprocess data from multiple sources to create a clean dataset for modeling WR performance based on athletic and draft metrics.

Pipeline Design & Details:

1. **Data Collection**: Loaded four datasets - combine.csv, draft.csv, players.csv, and yearly_player_data.csv.
2. **Filtering**: Selected only Wide Receivers (WRs) for analysis.
3. **Merging**: Merged datasets to bring together combine, draft, and performance using an inner join.
4. **Cleaning**: Removed redundant columns and replaced nulls values with median for the column. Missing values were addressed using the inner join above to make sure we have combine/draft data for every player we are analyzing.

5. **Normalization:** We chose not to normalize the data as it would hinder interpretability and decrease variation within the dataset.

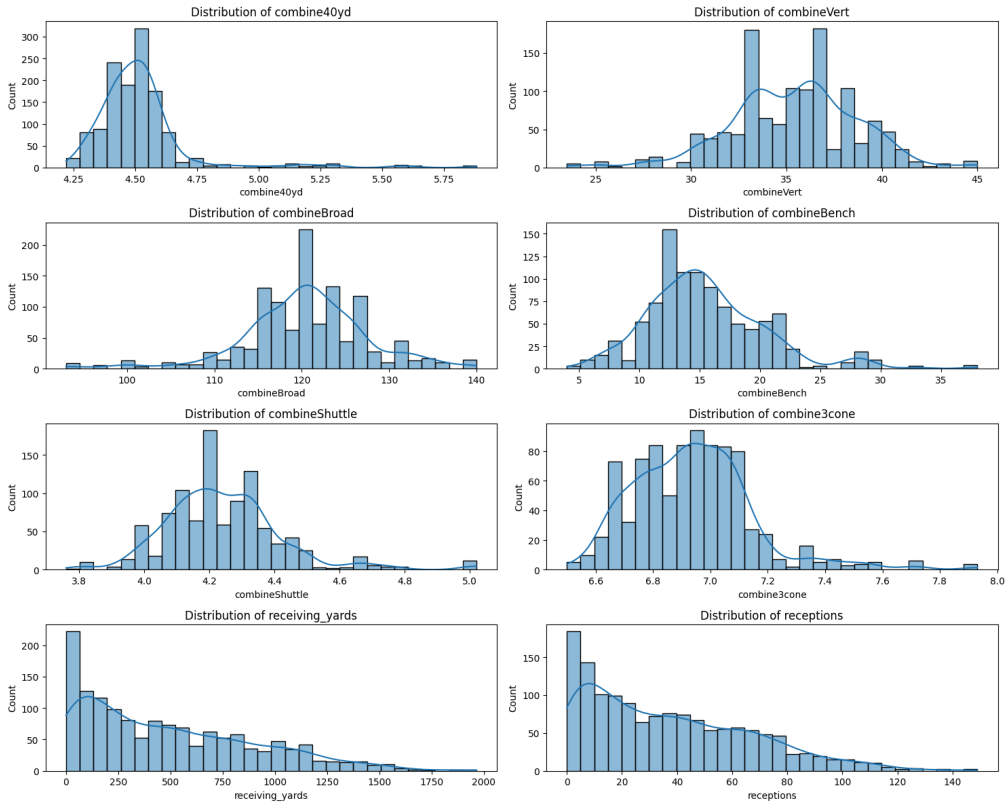
Output: A structured dataset containing WR combine metrics, draft round/pick, yearly stats, as well as overall career stats.

Exploratory Data Analysis (EDA)

Types of Analysis Used:

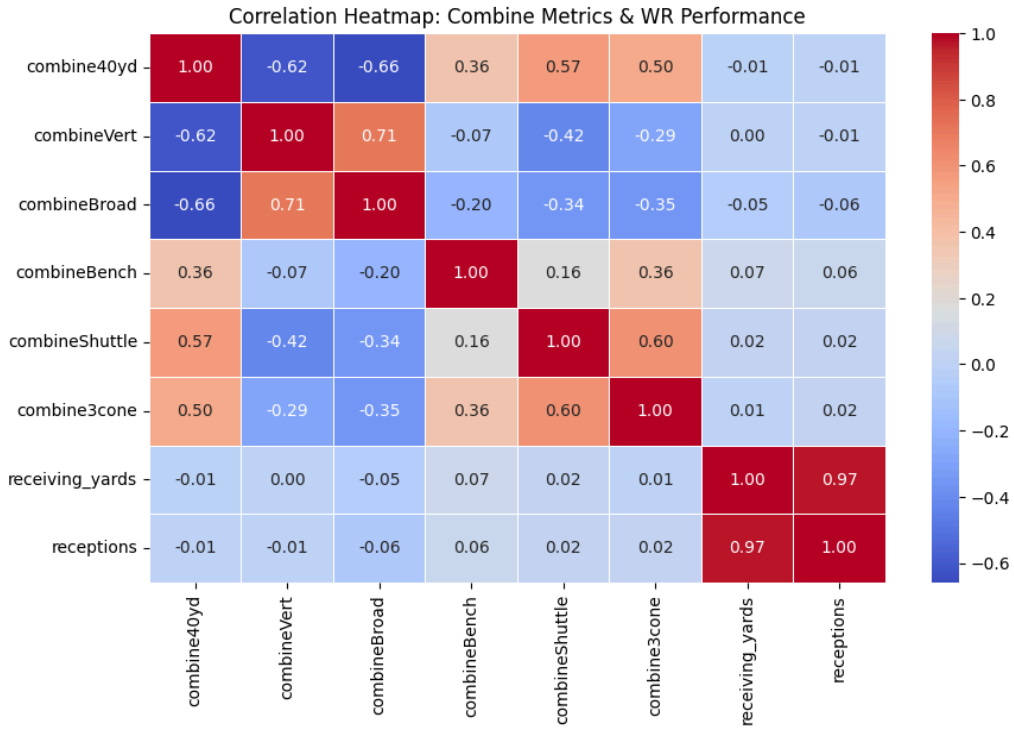
- **Univariate Analysis:** Histograms to examine distributions of combine metrics (40-yard dash, vertical jump, bench press, etc.).
- **Multivariate Analysis:** Correlation heatmap to identify relationships between combine metrics and WR receptions/receiving yards in a given season. Additionally, scatterplots were used to observe correlation between different combine metrics and fantasy points scored in a season.
- **Outlier Detection:** Boxplots to check for extreme values in combine stats and performance metrics.

Graphs:



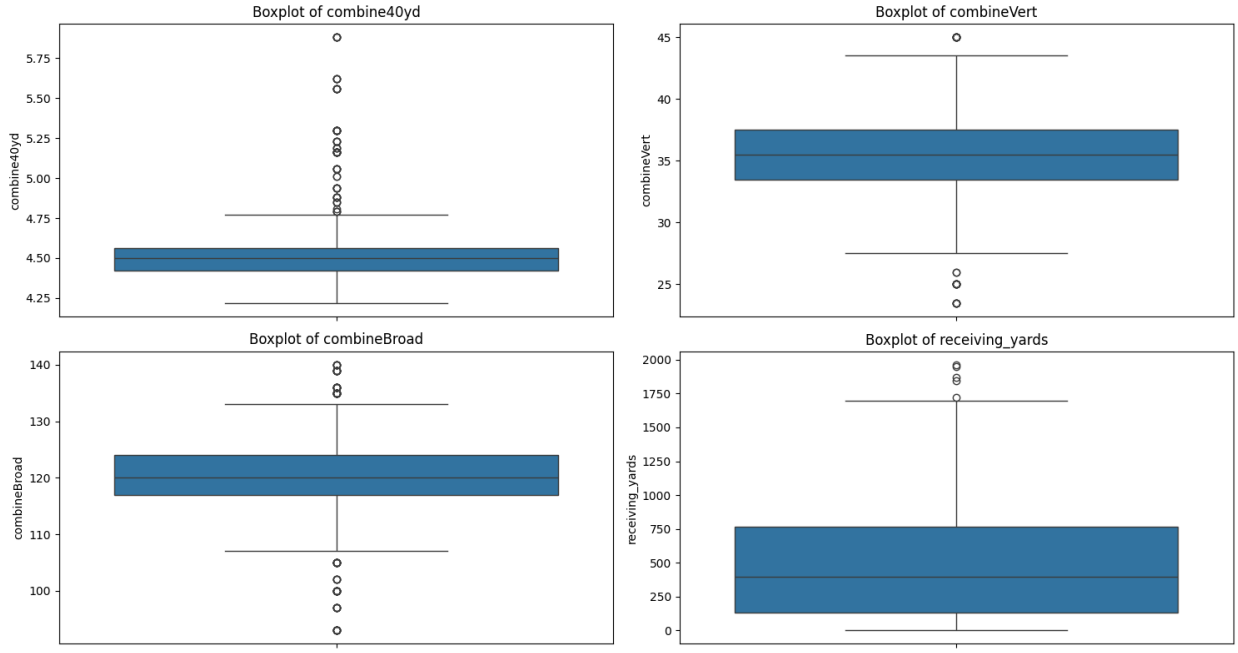
From the above histograms, the 40-yard dash times are concentrated around 4.5 seconds, with a few instances of much slower players in the 5+ second range. In contrast, broad jump results are left-skewed, peaking at around 120 inches. The shuttle run times center around 4.2 seconds, with a few outliers at round 5 seconds. Vertical jump measurements display a wide range, with multiple peaks between 30 to 40 inches, highlighting variability in players' explosive power. Bench press repetitions also show a bell-shaped distribution skewed slightly to the right, with most athletes achieving between 15 reps. The three-cone drill times, which measure agility and quick directional changes, have a nearly uniform distribution from 6.7 to 7.1 seconds, with a slight skew to the right.

We also observe that both receiving yards and receptions are heavily right-skewed. Most players accumulate fewer than 500 yards and 40 receptions respectively, but the distributions have long tails where a few players achieve significantly higher values. This indicates that while the majority of players perform at moderate levels, we see a fair number of players that reach above and beyond and perform significantly better than everyone else.

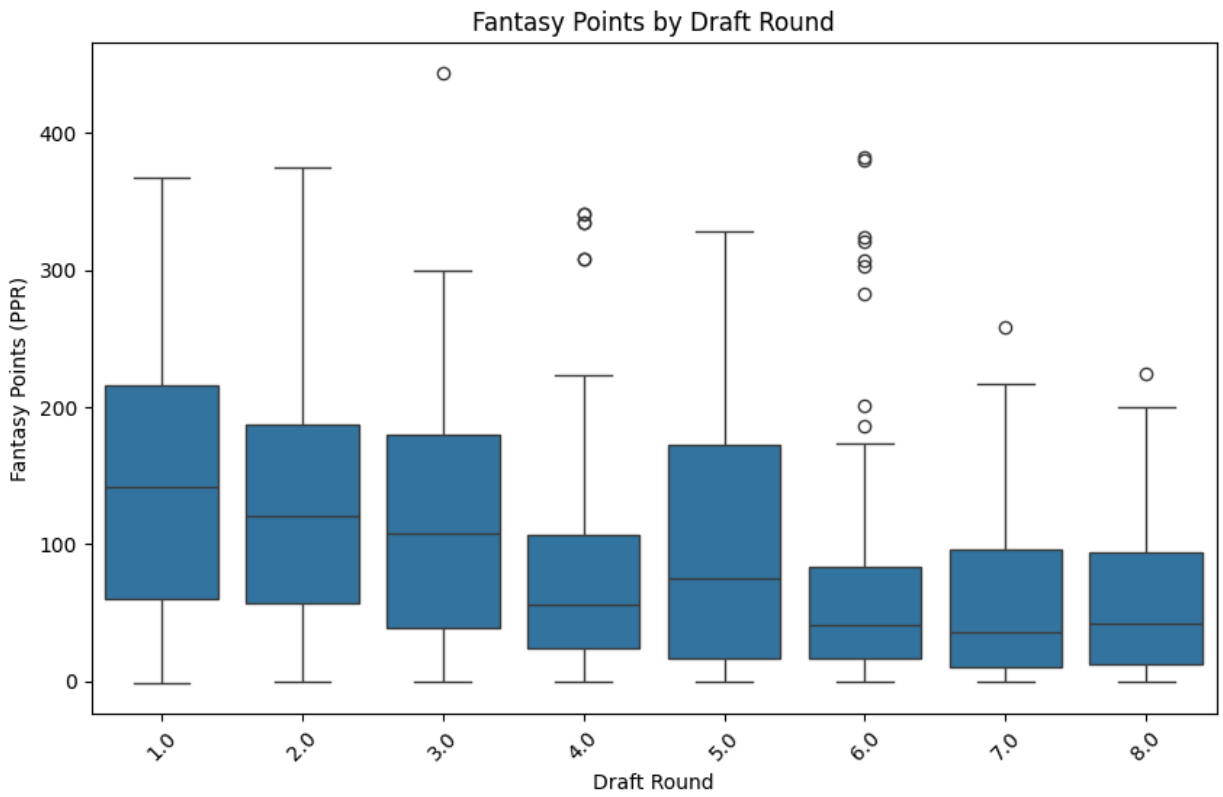


From the correlation heatmap, we see several notable relationships between metrics. Firstly, there is a strong correlation (0.71) between vertical jump and broad jump. Players with more explosive lower bodies are generally performing well in both events. Additionally, there is a moderately strong correlation (0.60) between shuttle run and three-cone drill. Both drills measure agility and quickness, suggesting that athletes who are agile in linear runs are also agile in drills requiring changes of direction. Furthermore, the 40-yard dash shows moderate positive correlations with the shuttle run (0.57) and three-cone drill (0.50). Faster sprint times tend to coincide with better performances in agility-related drills, which might be due to the

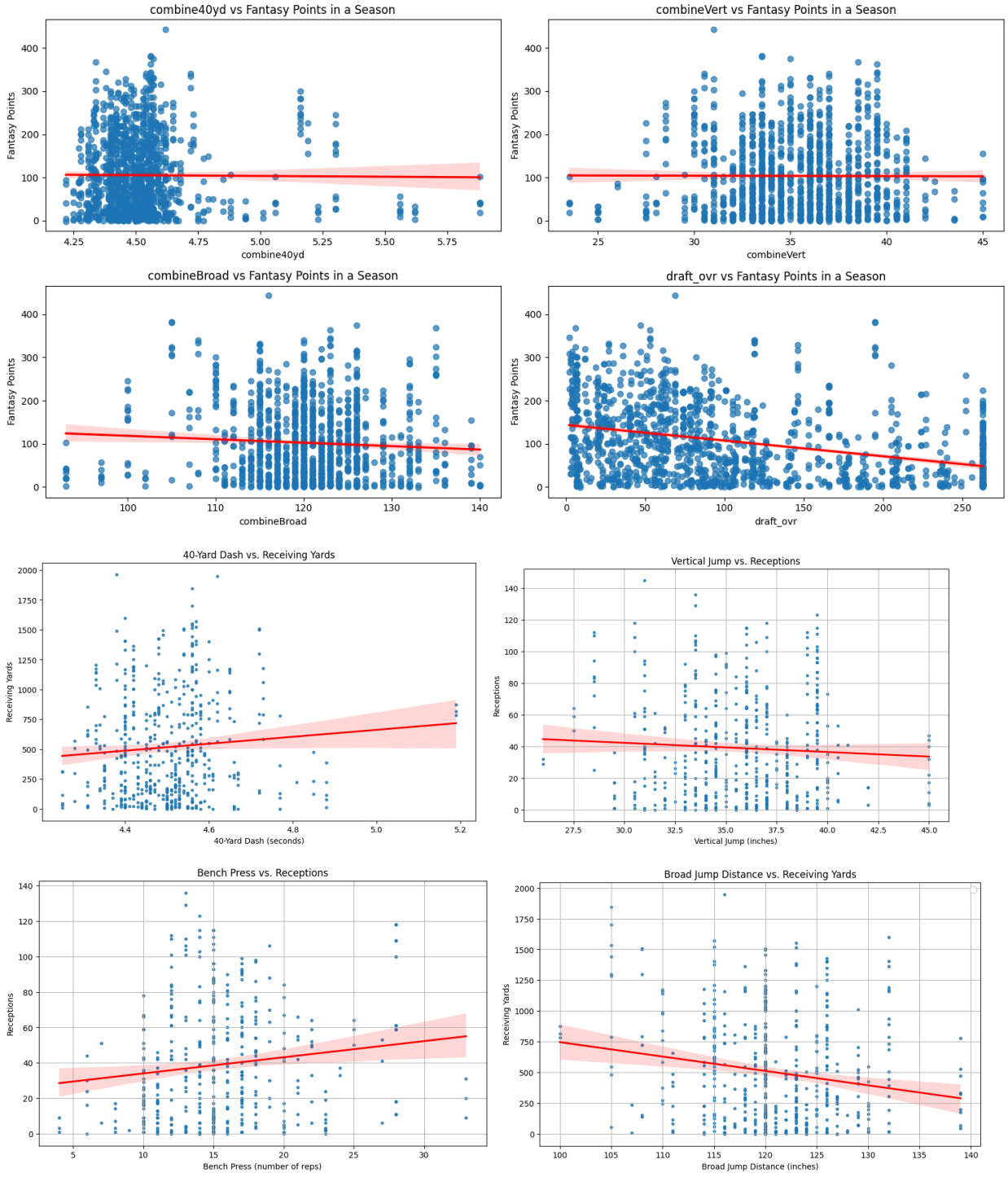
athletes' overall speed and reaction capabilities. Finally, the bench press (combineBench) shows very weak correlations with both receiving yards (0.07) and receptions (0.06). This suggests that upper body strength does not significantly predict receiving performance in games. It is also interesting to note that receiving yards and receptions have a very low correlation to each of the combine metrics, indicating that there are many different factors that lead to a Wide Receiver's success in the NFL.



These four plots show the distributions of the 40 yard dash, vertical and broad jump, and receiving yards. The 40-yard dash boxplot displays a few outliers on the higher end, and the distribution of receiving yards is skewed toward the higher end, reinforcing the histogram plots for these metrics.



From this figure, we see that players drafted in the first round typically score higher fantasy points, with a higher media. There is a noticeable decrease in both the median and the range of fantasy points from the first to the third round. This suggests that players drafted earlier are more likely to yield higher fantasy points, supporting the general strategy of prioritizing earlier draft picks for fantasy considerations. After the fifth round, the medians stabilize and the boxes become shorter, suggesting a decrease in variability of fantasy points. Although the ceiling for performance (as indicated by the whiskers and outliers) lowers, players drafted in these rounds tend to have a more predictable range of outcomes. It is interesting to point out that the general trend of points continues to decrease per round, except for a noticeable spike in the fifth round.



From these plots, we see that there doesn't seem to be much of a linear trend between the 40 yard dash and fantasy points in a season, nor the combine vertical or broad jump. However, we do notice that draft round has a bit of a relationship with fantasy points (as described in the previous visualization). An increase in bench press capability also seems to have a relationship with more receptions, as well as 40-yard dash and receiving yards and broad jump distance and receiving yards.

Feature Engineering

We plan to use regularization to identify the most impactful features for predicting fantasy scores. Specifically, we will apply Elastic Net regression, which balances L1 and L2 regularization, allowing us to both shrink and select features effectively. This approach will help determine which relevant features contribute the most to the model's predictions. Additionally, we will explore Principal Component Analysis (PCA) to reduce the number of features and enhance interpretability, ensuring the model remains both effective and easy to analyze. To handle missing values, we have replaced nulls with the median values of their respective columns, ensuring a more robust dataset for modeling.

Modeling Plan

Our modeling plan includes a diverse set of approaches to evaluate the effectiveness of different algorithms in predicting fantasy scores. We will implement Random Forest Regression, which is well-suited for handling non-linearity and provides valuable insights through feature importance analysis. Additionally, we will explore Gradient Boosting methods such as XGBoost and LightGBM, which are highly effective for structured data and can naturally handle missing values. As a baseline, we will use Multiple Linear Regression to establish interpretability and assess the impact of more complex models. To further explore non-linear relationships, we will experiment with a neural network, testing the effectiveness of deep learning on tabular data.

Progress Report

Completed So Far:

- Data cleaning and merging (Combine + Draft + Yearly Performance Data).
- Exploratory Data Analysis (EDA): Histograms, correlations, outlier detection, and analysis of plots.
- Identified key features for modeling.

Next Steps:

- Train initial models and evaluate performance.
- Handle missing values in NFL combine metrics.
- Optimize feature selection for predictive accuracy.

Team Contributions

- **[Vitush]**: Exploratory Data Analysis, generating visualizations (1-3).
- **[Praveen]**: Exploratory Data Analysis for Visualizations (4-6)
- **[William]**: Exploratory Data Analysis, generating remaining visualizations, analysis for visualizations.
- **[All]**: Researching modeling techniques, setting up training pipeline.

Risks and Mitigation

Risk 1: Small dataset size after filtering to WRs with both combine and performance data

- **Mitigation:** Consider feature augmentation.

Risk 2: Model overfitting due to high-dimensional data.

- **Mitigation:** Use cross-validation, feature selection, and regularization techniques.

References

1. Burke, B. (2020). "How Combine Metrics Predict NFL Success."
2. PFF (2021). "Athletic Testing & NFL Draft Outcomes."
3. https://www.profootballnetwork.com/what-is-ras-explaining-athletic-testing-metric/
4. https://www.kaggle.com/datasets/philiphyde1/nfl-stats-1999-2022?resource=download
5. Kaggle Dataset: "NFL Player Statistics Dataset (2004-Present)" - https://www.kaggle.com/datasets/toddsteussie/nfl-play-statistics-dataset-2004-to-present

---