**Databases and Sources**

**Combine Data (source 1):**

[https://nflcombineresults.com/nflcombinedata\_expanded.php?year=all&pos=QB&college](https://nflcombineresults.com/nflcombinedata_expanded.php?year=all&pos=QB&college=)**Combine Data (source 2):**

<https://www.kaggle.com/datasets/savvastj/nfl-combine-data>

**Career Data:**

<https://stathead.com/football/>

<https://stathead.com/football/player-season-finder.cgi>

**Career Data (source 2):**

<https://www.nfl.com/stats/player-stats/category/receiving/2022/REG/all/receivingyards/DESC>

**Behavior Driven Development (BDD) info:**

<https://cucumber.io/docs/bdd/>

<https://behave.readthedocs.io/en/latest/gherkin/#step-data>

**---------------------------------------------------------------------------------------------------------------------**

**Project Proposal**

Brought to you by Meriam Moujahed, Arianwyn Frank, Jinfeng Chen, Harrison Keyser

**Topic Domain**

We want to predict the career performance and lifespan of professional athletes, specifically focusing on football players, using a Multivariable Regression and Correlation Matrix analysis. We believe that a player's NFL Combine performance metrics, such as their speed, strength, agility, and endurance, can provide valuable insights into their performance and longevity in the sport. By collecting and analyzing historical data on player combine performances and their career trajectories, we intend to develop a predictive model that can estimate the potential duration and performance of a player's career.

**Research Question and Hypothesis**

Does the NFL Combine help predict the quality of an NFL prospect? Hypothesis: the combine will help predict the career statistics but not the career lifespan.

**Feature 1: Predict position-specific career statistics based on the combine performance**

The output features of the algorithms will be various career performance data, such as career stats (touchdowns, points scored, receiving yards, etc).

**Scenario 1: Input combine stats**

* **Given** a dataset of combine statistics for 10 players,
* **When** the data is loaded into the system,
* **Then** ensure that the resulting data frame contains less than 5% NA (missing) values.

**Scenario 2: Create regression algorithm**

* **Given** a dataset of combine statistics for 10 players,
* **When** the regression algorithm is applied to the dataset,
* **Then** generate a plot that visually represents the relationship between combine scores and career performance metrics, and provide the equation of the regression line.

**Scenario 3: Create correlation matrix**

* **Given** a dataset of combine statistics and player ages,
* **When** the correlation matrix is generated,
* **Then** produce a report that includes the correlation coefficient between each exercise (combine stat) and age. For example, "Bench press has a correlation of 0.03 with age."

**Feature 2: Predict player career lifespan based on the combine performance**

Multivariable Regression and correlation matrix. The output model could predict various career longevity data, such as age of retirement, seasons played, games played, and injury frequency.

**Scenario 1: Input Combine Stats**

* **Given**: Combine statistics for multiple players.
* **When**: The system receives this data.
* **Then**: Ensure the data has x number of rows

**Scenario 2: Create Regression Algorithm**

* **Given**: Cleaned combine statistics.
* **When**: The system initiates the regression algorithm.
* **Then:** Verify that the algorithm produces a realistic correlation equation (aka generates \*something\* equation adjacent)

**Scenario 3: Generate a Plot with Equation (for a dataset of 10 players)**

* **Given:** The regression algorithm is executed.
* **When**: The system receives data for 10 players' combine scores.
* **Then**: Ensure the system generates a plot that displays the relationship between combine scores and career lifespan, along with the regression equation.

**Scenario 4: Create Correlation Matrix**

* **Given**: Cleaned combine statistics.
* **When:** The system initiates the correlation matrix calculation.
* **Then:** Verify that the correlation coefficients between each exercise and age are computed correctly (hand calculate to verify)

**Scenario 5: Output Correlation Coefficients**

* **Given**: The correlation matrix is calculated.
* **When:** The system is asked to provide the correlation coefficients.
* **Then:** Ensure the system outputs the correlation coefficients for each exercise's relationship with age (e.g., "Bench press has a 0.03 correlation with age").