NFL Attributes:

What is important?

Each year, NFL teams spend millions of dollars signing players to contracts without knowing if the player will underperform or outperform the contract. Will the player who had an amazing season last year have another one this year? Or did he just try extra hard the year before so he could get a big contract and now he intends to coast for the rest of the year? There is also a scouting event for select college players where teams record the potential players speed, strength, and agility. Are there any positions where a team just takes the player with the best result in an event? This project is designed for NFL teams that are trying to decide what makes a good NFL player. By using linear regression, I hope to find what attributes are most important for each position and then create models for each position so teams can know what players to sign and which ones to avoid.

**Data Wrangling**

Cleaning Steps

To find the data, I went to Pro Football Reference and found spreadsheets with NFL combine statistics. I also found csv files from the last four years with the Madden franchise player ratings. The data was not uniform, especially in the format of player names. To fix this problem, I replaced the commas and slashes with spaces and broke each string into a first name and a last name. I then merged the first name and last name column so every players’ name was in the same format.

Once I had all the data cleaned and together, I concatenated all four years of Madden data vertically and then removed all duplicate entries by name after the first occurrence to prevent multicollinearity problems while doing future analysis. This also created one instance of each player and was their most recent season. After doing this, I merged the Madden data with the combine data using a left merge so only the combine numbers with corresponding Madden players were merged.

I also made a separate spreadsheet that had of the data merged by player name. This data included the college each player attended and was used for the data storytelling section.

Missing Values

Missing data came from players who did not participate in the combine. To replace these values, I broke down all the players into position groups and filled each missing value with the position average for each column.

data is not as detailed as it could be. For instance, for this dataset, there is only one category for run blocking. That category could be broken down into the components that make up the run blocking skill so more detailed data could be generated.

Inferential Statistics

Typically, when teams are evaluating quarterbacks, they look for big quarterbacks that can throw the ball far and run. Based on the numbers though, the most important attribute a quarterback can have after awareness is mid-range passing. These passes are typically 5 to 15 yards away and require excellent timing and accuracy. Awareness is a very difficult statistic to quantify especially on a 0- 100 scale. A quarterback’s ability to throw mid-range passes however is easily observable while watching them play a game. By using this metric, teams could get better data for drafting quarterbacks

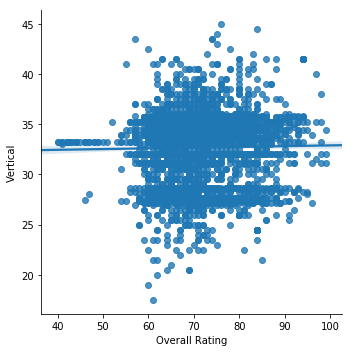
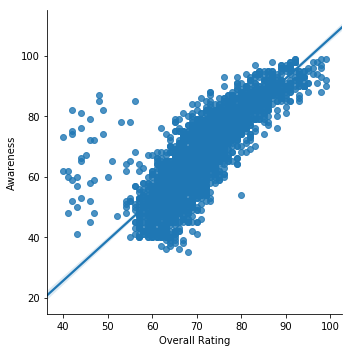
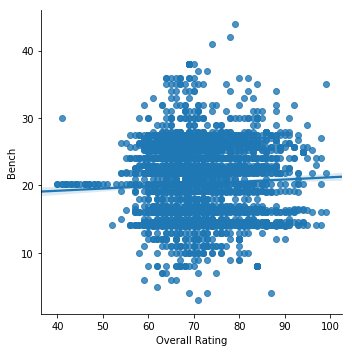
Combine Event Correlations

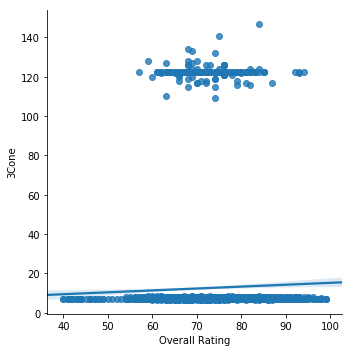
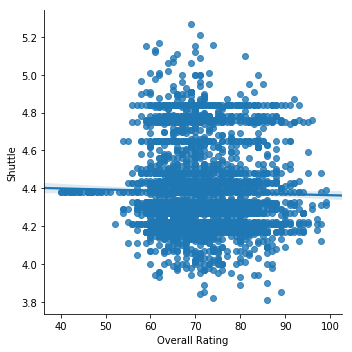
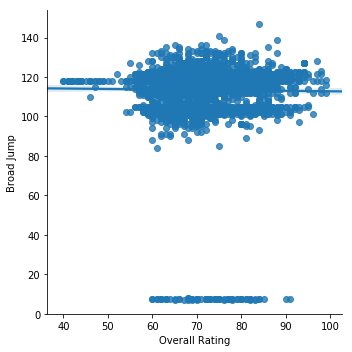
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Combine Events** | **Bench** | **Vertical** | **Broad Jump** | **3Cone** | **Shuttle** | **Height** | **Weight** |
| **Bench** | 1.0000 | -0.5411 | -0.3408 | -0.1432 | 0.6749 | -0.0114 | 0.6989 |
| **Vertical** | -0.5411 | 1.0000 | **0.4557** | 0.1848 | **-0.7854** | -0.0327 | -0.7364 |
| **Broad Jump** | -0.3408 | 0.4557 | 1.0000 | 0.1133 | -0.4107 | 0.0139 | -0.3557 |
| **3Cone** | -0.1432 | 0.1848 | 0.1133 | 1.0000 | -0.1802 | -0.0563 | -0.1849 |
| **Shuttle** | **0.6749** | -0.7854 | -0.4107 | -0.1802 | 1.0000 | 0.0177 | 0.7847 |
| **Height** | -0.0114 | -0.0327 | 0.0139 | -0.0563 | 0.0177 | 1.0000 | 0.0359 |
| **Weight** | **0.6989** | **-0.7364** | -0.3557 | -0.1849 | 0.7847 | 0.0359 | 1.0000 |

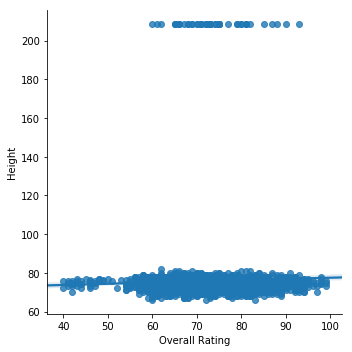
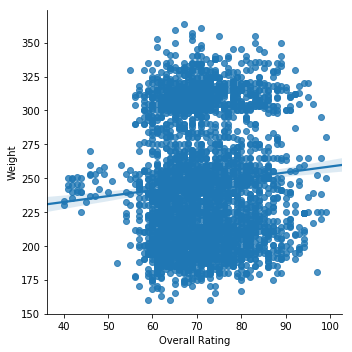
Above are the correlations for the combine events with each other. In general, the bigger the player is, the stronger they are. Conversely, the smaller a player is, the faster they are. It also looks like players who are fast can typically jump high and far as well.

**Awareness Correlation vs Combine Event Correlation**

Pictured from left to right: Bench Press, Awareness, Vertical, Broad Jump, Shuttle, 3 cones, Weight, Height







Pictured above are eight different scatter plots showing the seven different combine drills along with the most highly correlated attribute, awareness. As you can see, the physical attributes are essentially uncorrelated with Overall Rating. While there are certainly athletic minimums for the NFL (I would not be able to play no matter how high my awareness was), it seems that once a player reaches a certain minimum level of athleticism, it quickly drops in importance and other, non-physical attributes become more important.

**Regression**

Once all the data was sorted and organized, I created a linear regression model. The first step in creating the model was to take all the variables with a correlation greater than 0.5 or less than -0.5. The next step was to check for multicollinearity. I did this by checking VIF scores and removing the highest score until all scores were below 5. I also checked p-values and made sure each variable was below 0.05. Next, I checked for heteroskedasticity and there was not any significant amount found in the Breusch-Pagan test results.