Matt Tilley

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 had statistics for every season played by every player in the Madden dataframe over the last four years. I created this data frame to create 2, 3 and 4 year averages of each player tocreate a variable to account for players who were better and played longer.

Column Creation

Combine Ratios

Each combine event measures a different attribute, so I thought it would be interesting to see if different ratios were more strongly correlated. Does a player with a high bench press and a low vertical differ from players with a low bench press and a high vertical?

Averages

To account for players who played the most years (and would be the best players), I created a different data frame that merged all the columns horizontally by player name. I then created separate columns with a four, three, and two-year average.

Missing Values

When doing this analysis, there were two different types of missing values. The first was from the Madden data and existed because not all players played all four years. For this data, I simply replaced all of the missing data with 0’s because this would cause lower averages in the columns this data was needed for.

The second type of missing data was 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 mean for each column. I wanted to use position means instead of overall means because there is such a wide difference in the size of players on the data set.