Statistical Data Analysis

Once we accomplished the exploratory phase of our project, we moved on to do more rigorous statistical analysis on our data, using the techniques learnt in the corresponding section of the curriculum.

Most of our previous analysis consisted in highlighting general trends, especially between college football and professional league data. Of course, most of the cfb stats positively correlated with the nfl ones - those that relate to receiving stats particularly do - while the combine numbers correlate more moderately; but those are relative observations.

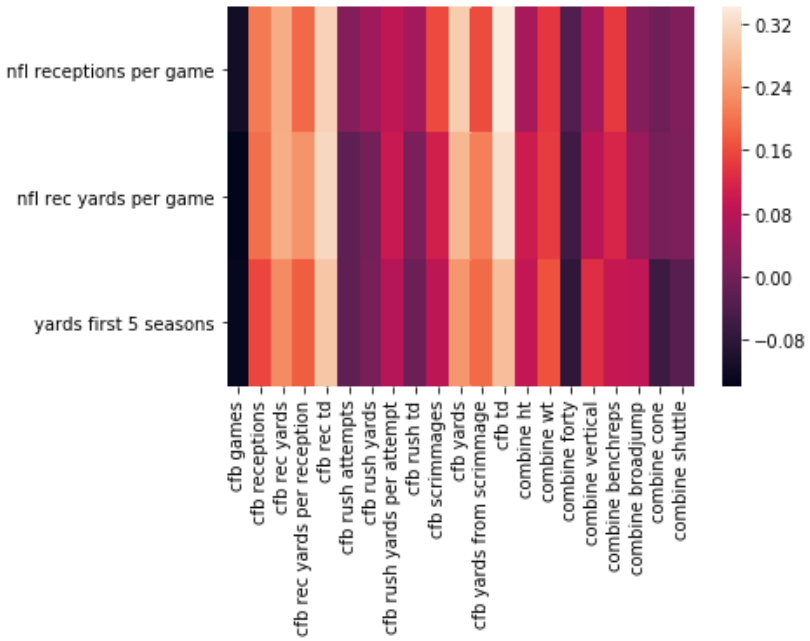
We now decide to pick 3 potential targets (X_i) that correlate the most with the input data (Y_j). Using the correlation matrix between our numerical data, we pick targets with the highest aggregate correlation factor:

$$C_i = \sum_j corr(X_i, Y_j)$$

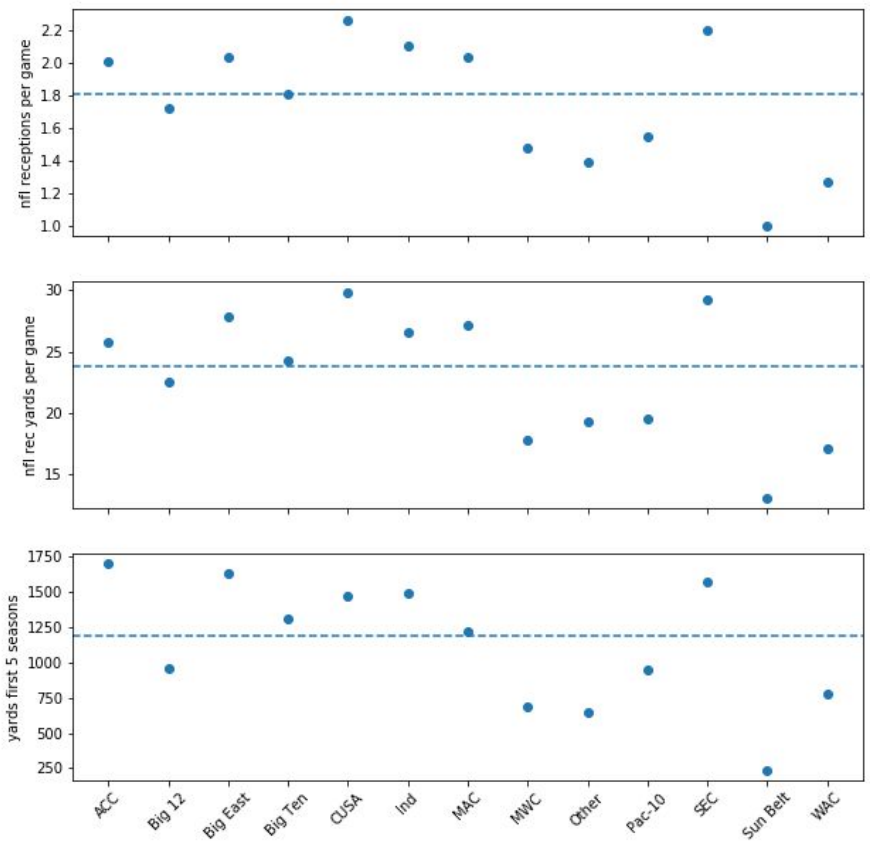
We find the following targets:

- **Nfl receptions per game**
- **Nfl receiving yards per game**
- **Yards during the first 5 seasons.**

We plot the correlation matrix of those 3 targets through a heatmap:



and also account for non-numerical data such as cfb conference of origin of the player when drafted:



We now design statistical tests that will highlight the presence or absence of correlation between an input variable and a target. We do so by considering or defining subgroups in our data. We chose our level of confidence to be $\alpha = 95\%$, compute the p-value for each test and reject (or not) the respective null hypothesis.

Test1: cfb conference / SEC vs Big12

We run a bootstrap inference on those two categories and we do it with respect to the mean of each of our 3 metrics X_i. The null hypothesis can be stated as “The sec and big12 samples are from the same distribution with respect to the mean of X_i”. We found the following p-values:

- NFL receptions / game: p = 0.0629
- NFL rec yards / game: p = 0.0637
- Yards first 5 seasons: p = 0.0213

Test2: cfb rush attempts

Group1 and group2 are the subgroups of players with cfb rush attempts respectively below and above the median. This time, we test for the independence of the two samples w.r.t the 3 metrics using a scipytest:

- NFL receptions / game: p = 0.2450
- NFL rec yards / game: p = 0.3312
- Yards first 5 seasons: p = 0.3569

Test3: cfb receiving touchdown

Group1 and group2 are the subgroups of players with cfb receiving touchdowns respectively below and above the mean. We follow the same procedure as test2 and find:

- NFL receptions / game: p = 0.00000127
- NFL rec yards / game: p = 0.00000012
- Yards first 5 seasons: p = 0.00000060

The confidence interval we chose was [0.025, 0.975]. Test1 and test3 return on average relatively small p-values that fall outside that interval, which indicates that under the stated null hypothesis, the probability of returning a result as extreme is very small, meaning we can dismiss the hypothesis. This confirms what we saw on the correlation heatmap and the plot. Test2 on the contrary returns large p-values, meaning the null hypothesis is likely to be correct, once again confirming what we saw on the correlation heatmap.