Capstone #1 Project Milestone Report

Pawandeep Jandir

Introduction

The Problem

Professional sports are a lucrative and competitive industry. The National Football League (NFL) is the biggest sports league in the world. As is the norm in American professional sports, amateur players are selected via a league-wide draft, in this case the NFL Draft. A total of 256 selections are made over 7 rounds. In the lead-up to this event, the NFL holds the NFL Scouting Combine (Combine): a long-running showcase where incoming players perform physical and mental tasks for NFL teams. Examples include the 40-yard dash, vertical jump, 3 cone drill, Wonderlic (intelligence test), and personal interviews. While this is only a portion of the draft scouting activity, it has tended to give lots of weight to the overall process. This analysis will attempt to tease out any predictive power that exists between the quantitative Combine results and draft round. Lots of resources are spent scouting players and determining draft rankings. This project can help a team provide a better draft grade for players and can also be used to estimate when a particular player may be drafted so that a draft trade can be used to target that specific player.

The Client

The potential target audience are NFL teams. They would be given a prediction model which outputs the probability of which round a player would be drafted in. They can incorporate this into their draft rankings and strategies. This can allow a team to target certain players with greater certainty given their own assigned draft picks or traded draft picks.

The Data

Three data sources are used in this project. The first is from the NFL Savant website and has data from 1999 to 2015. This is not the only website with this information but it also has provides measurements for player hands and arms. In order to check for player and drill/task completeness, Sports Reference is used. The third source, DraftHistory, is needed for NFL Draft result completeness. We know that not all players who participate in the Combine get drafted. Conversely, not all players who get drafted participate in the Combine. Further, not all players who are at the Combine will participate in all drills and tasks. So it is expected that there are null values throughout the Combine datasets. This must be kept in mind throughout the entire process.

Potential player college data and statistics can also be used to perhaps add more predictive power to this analysis. Not all colleges produce equal amounts of draft picks. Some colleges are known powerhouses, churning out many players who end up being drafted, often quite high. Additionally, being on a good or high performing college team also boasts a players profile and chances in the NFL Draft. So adding this type of information could provide additional handles on predicting draft round.

Data wrangling

Acquistion

After identification of data sources, the next step is to obtain it somehow. The list below gives the details for each data source.

1. NFL Savant : This is by far the easiest of the trio. The site offers a csv file of the Combine data. This can then be read into a pandas dataframe very easily.

2. Sports Reference : The data stored by Sports Reference is in html tables. Thus, the site has to be scraped using requests and BeautifulSoup. After getting the full contents of the html table using those packages, we construct a function which can create a dataframe from it. After this dataframe is created, a single column consisting of draft team, round, pick, and year is split up into their own separate columns with str.split(). Each year is on a separate page, so we loop through all the webpages one at a time while appending each created data frame. At the end, the dataframes are concatenated to a single dataframe. To avoid making repeated calls to the Sports Reference website, this dataframe is saved (via pickle) locally. This should allow for better reproducibility as well.

3. DraftHistory : Similar to Sports Reference, Draft History also has their data available on their website in html tables. A very similar approach was used since much of the framework to scrape was already in place. The approach is identical up to the html table to pandas dataframe conversion. After that, there are differences in how the resultant dataframe should be formatted. The draft round column is littered with "\xa0" strings so that needs to be removed. Also the draft round column needs to be forward filled since the website html tables only displays the draft round number when there was a new round. Looping over the relevant years, this dataframe is also saved to a local pickle file for later analysis.

Cleaning

The largest portion of time was devoted to cleaning and ensuring data quality. The three datasets were worked in order. Immediately an issue was spotted when we read in the NFL Savant csv file. Two rows had elements which had commas in the field. Obviously this is a problem for csv files. However, the issue was confined only to two rows, so instead of trying to find a general method to deal with such edge cases, we went and manually changed the csv file to remove the offending commas. This way, we could load up the data and continue.

Considerable time was spent looking at what data was missing and how best to fill it (if at all). One way already mentioned was using additional datasets. We loaded up the Sports Reference dataframe and ensured the columns are the correct type. However, it is obvious this dataset is also not complete. We used pd.merge() to full outer join the two dataframes and include an indicator variable for manual inspection of the joined dataframe. We wrote a small function which compared and combined the values of two columns (one from each dataset) to see the percentage of null values. This way there is a quantitative way to determine how combing the two datasets (for a particular column) can recover missing data. On average, there is approximately a 10% recovery. This works about how we expected. Unnecessary columns are dropped.

At this point it is worth cleaning up the player college values. Because we joined two different data sources not every college is listed with the same name. For instance, the University of Southern California can be referred to as Southern California or just USC. This needs be consistent across whatever datasources we use. After this step, we turned our focus on the draft round and pick data which seems to be missing more values than expected. A closer look reveals the 2014 year is totally missing these values. Even worse, both data sources mislabel draft round completely. In both they refer to the pick within a round instead of the round itself, while the pick refers to the overall pick and not the pick within a round. This necessitates inclusion of the third datasource, DraftHistory, to get the correct draft round and pick data.

This time we used a left join to merge the two dataframes. This is because not all drafted players participate in the Combine. Any players missing in the DraftHistory dataset mean they were not drafted, and thus have legitimate null values in their draft pick data. This third dataset is also used to fill the draft team column which seemed to be missing some values. After combining columns as before, a more complete dataframe is now forming.

The last step is to fill in the null data throughout the dataset. There are various ways to do this and we explored the two basic ways: mean and median imputation per column. A quick look at the outliers and statistics in each column reveal no particularly significant differences between the two. However, there are more outliers than expected, so to be safe, we imputed the null values with the relevant column median and not the mean. The outliers, perceived or otherwise, will not be changed at this time since these are real measurements. The last missing values to change are the draft round and pick data. For now, those values are set to -1, though this may change later.

Among the many columns in the datasets we have combined, only one was added. We grouped similar player positions together into a positiongroups column. This is because positions can be fluid going into the NFL so it is very useful to view college players at this coarser level. Additional columns may also be added relating to a player's college stats and a player's college itself in the future.

Columns

At the end of this process we have many variables or columns in our dataset. The table below lists all of these columns with descriptions and their potential in the prediction model. Note the twelve columns marked with an asterisk, *, are defined as the variables of interest which will be discussed later in this document.

Column	Description	Notes	Useful in prediction?
name	Name of the player	-	No
year	Draft year	-	Possibly
college	Player's college attended	-	Possibly
position	Player's college position	Scraped value	Yes
positiongroup	Player's college position group	Assigned value	Yes
height*	Player's height (inches)	-	Yes
weight*	Player's weight (pounds)	-	Yes
fortyyd*	Player's 40-yard dash (seconds)	Full description	Yes
vertical*	Player's standing vertical jump (inches)	Full description	Yes
bench*	Player's bench press of 225 pounds (reps)	Full description	Yes
threecone*	Player's 3 cone drill (seconds)	Full description	Yes
shuttle*	Player's shuttle run (seconds)	Full description	Yes
broad*	Player's standing long jump (inches)	Full description	Yes
wonderlic*	Player's Wonderlic test result	Sparsely populated	Possibly
nflgrade*	Player's NFL Grade determined by experts	Sparsely populated	Possibly
arms*	Player's arm length (inches)	-	Yes
hands*	Player's hand length (inches)	-	Yes
team	Team that drafts player	-	Possibly
round	Player's draft round	Target variable	N/A
pick	Player's draft pick number within a round	-	Possibly
overall	Player's overall draft number	-	Possibly

Data Exploration

Initial and exploratory data analysis was performed using various visuals and inferential statistics. The purpose was to gain some insight into how draft round is influenced by all the other variables in the cleaned dataset. Because quarterbacks are generally thought of as the most important player on the team, they are given outsized importance in the draft. For that reason, we focused on quarterbacks.

Visual Analysis

One of the first things to ask is how many quarterbacks are in the dataset per year. Is it pretty consistent over the years or does it vary? We can plot this, as shown in Fig. 1. Each year has the number of quarterbacks printed at each point. The mean is represented by the red dotted line.

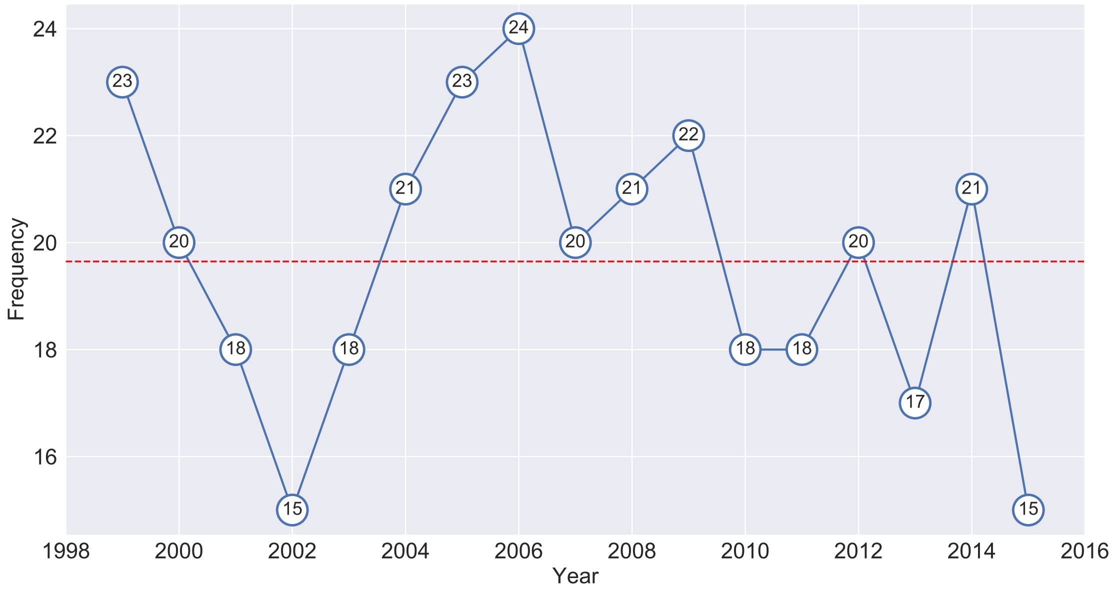

Figure 1: Quarterbacks per year.

We can see that generally there 20 or so quarterbacks every year with some notable exceptions. Along with a few years with a above average number of quarterbacks, like 2006, there are a few years with a below average number of quarterbacks, like 2002. There are probably a multitude of factors which we can hazard a guess for, but it is likely outside the scope of this project. We can also look at the distribution of quarterbacks as a function of draft round. Remember that a round of -1 means the player was undrafted. Shown below is the table of quarterback count per round.

Draft Round	-1	1	2	3	4	5	6	7
Count	141	47	19	22	23	29	27	26

It should not be too surprising to see that a large portion of quarterbacks do not get drafted. The quarterback is the most important position on the team so it is also the most scrutinized. There are only a few quarterbacks on any single team, so teams are judicious in selecting potential replacements. In that same vein, it should be expected that the first round is the most popular to draft a quarterback. It is the first opportunity to select a player and quarterback-needy teams likely need to prioritize that position over others. This importance means teams often gamble to get “their” guy who can be the starting quarterback for a decade or more. It is is also worth checking out how quarterbacks are drafted per round per year. This is best visualized in a heatmap, as shown in Fig. 2 below.

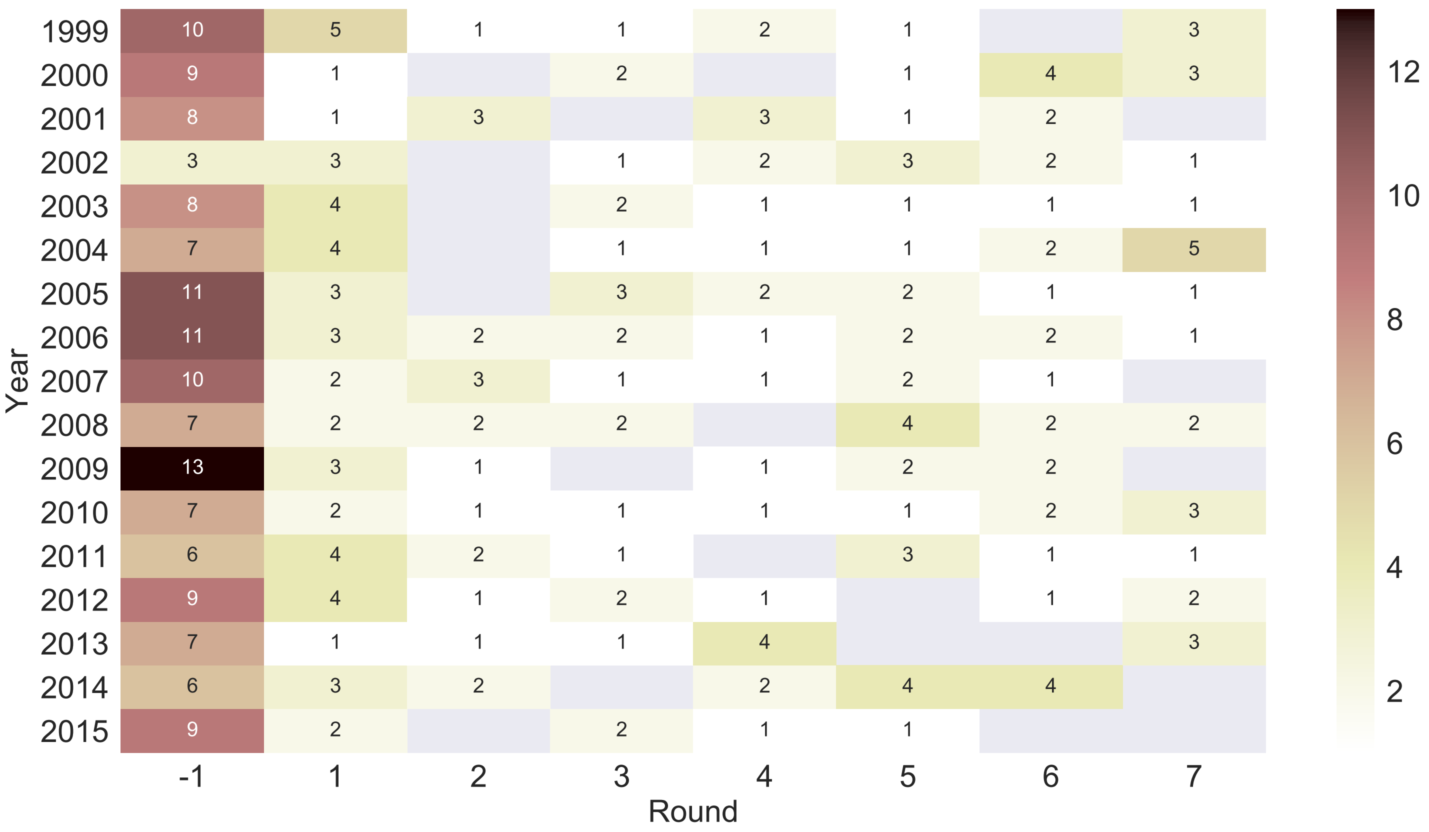

Figure 2: Quarterbacks per round and year.

We can look at this plot along with the previous plot to note some interesting years. In general, a quarterback is always selected in the first round. Again, no surprises here. The 2009 draft is quite surprising. It had an above average 22 quarterbacks but also let 13 of them go undrafted. Yikes. Only 2015 had a (marginally) worse percentage of players go undrafted.

We can also look at which teams have drafted quarterbacks by plotting it. This is done in Fig. 3 below. The red line represents the median of the number of quarterbacks drafted per team.

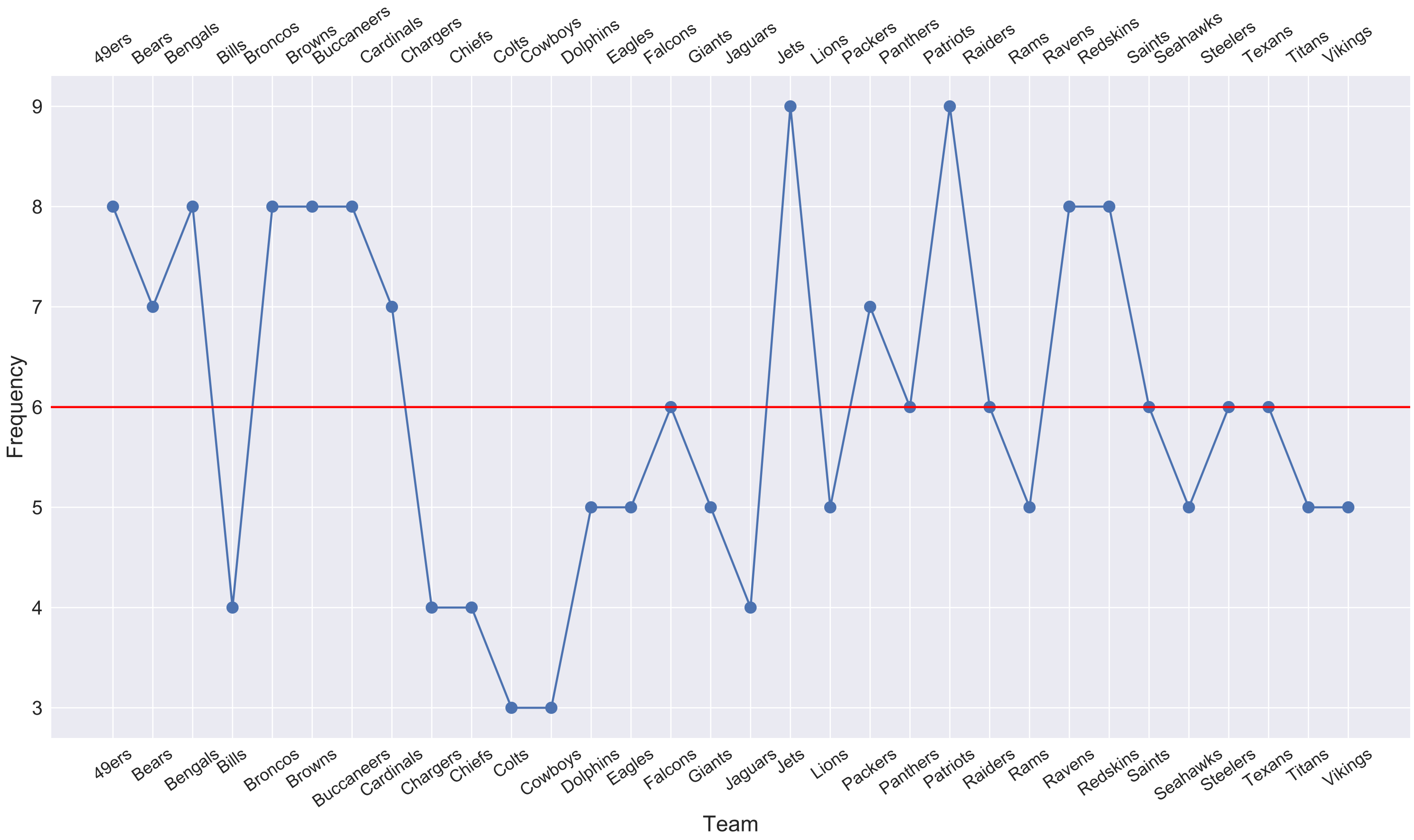

Figure 3: Quarterbacks drafted per team.

Considering this dataset covers 1999 to 2015, the teams with the most drafted quarterbacks make some sense (given knowledge of the NFL landscape). The teams who have tended not to have a long term starter at quarterback during this period show peaks in this distribution, although a notable exception does exist with the Patriots.

Let's switch gears to some of the actual measurables in the data along with the draft round information (the target variable in the analysis). We can pick out those interesting columns, as mentioned earlier, and make a heatmap as in Fig. 4. This shows an overall picture of how the variables are correlated. Positive and negative values are in purple and green, respectively.

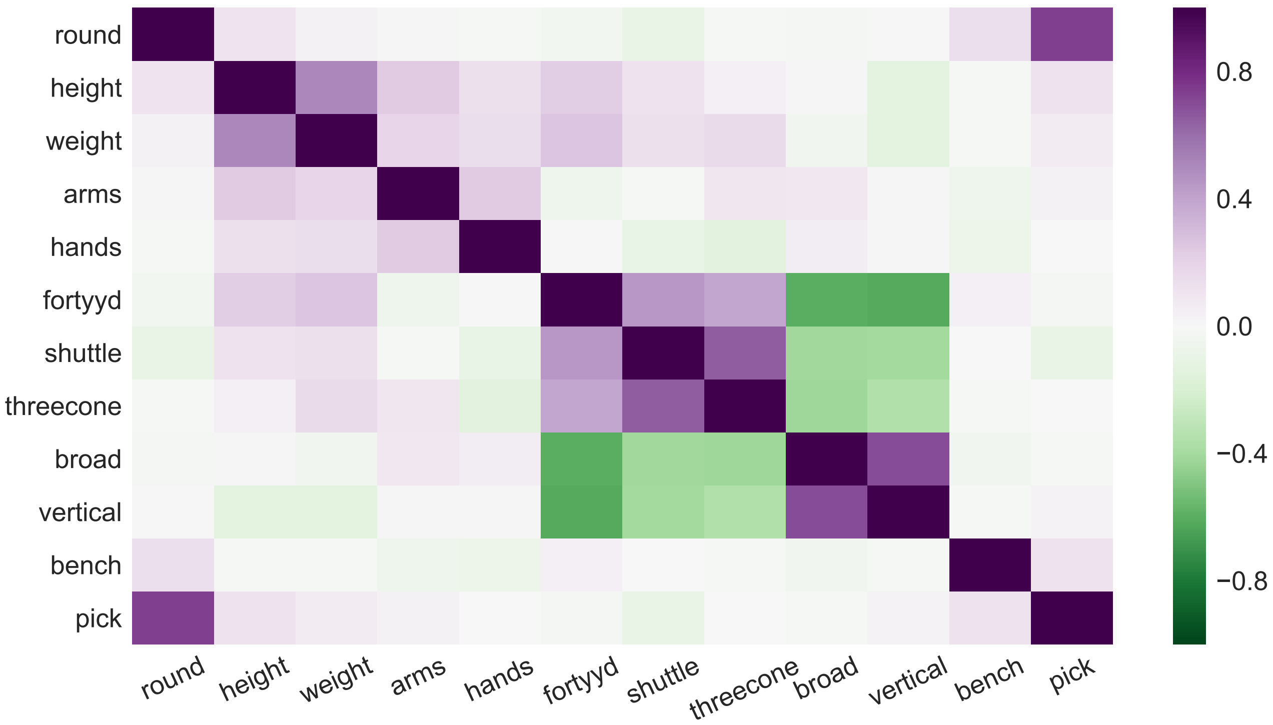

Figure 4: Correlations for all quarterbacks.

We can see how certain variables are pretty well (anti-)correlated. For instance, the fortyyd and vertical are relatively strongly anti-correlated. These values represent the 40 yd sprint and vertical jump drills, so perhaps this should be expected: the muscle groups involved are different for each drill. This is reinforced by the strong correlation between the broad (or standing long jump) and vertical jumps. However, this data is for all quarterbacks in this dataset. How does it look for quarterbacks that were actually drafted? We can re-do this heatmap as in Fig. 5 with that constraint. As an added visual aid for this plot, we can also print the correlation values.

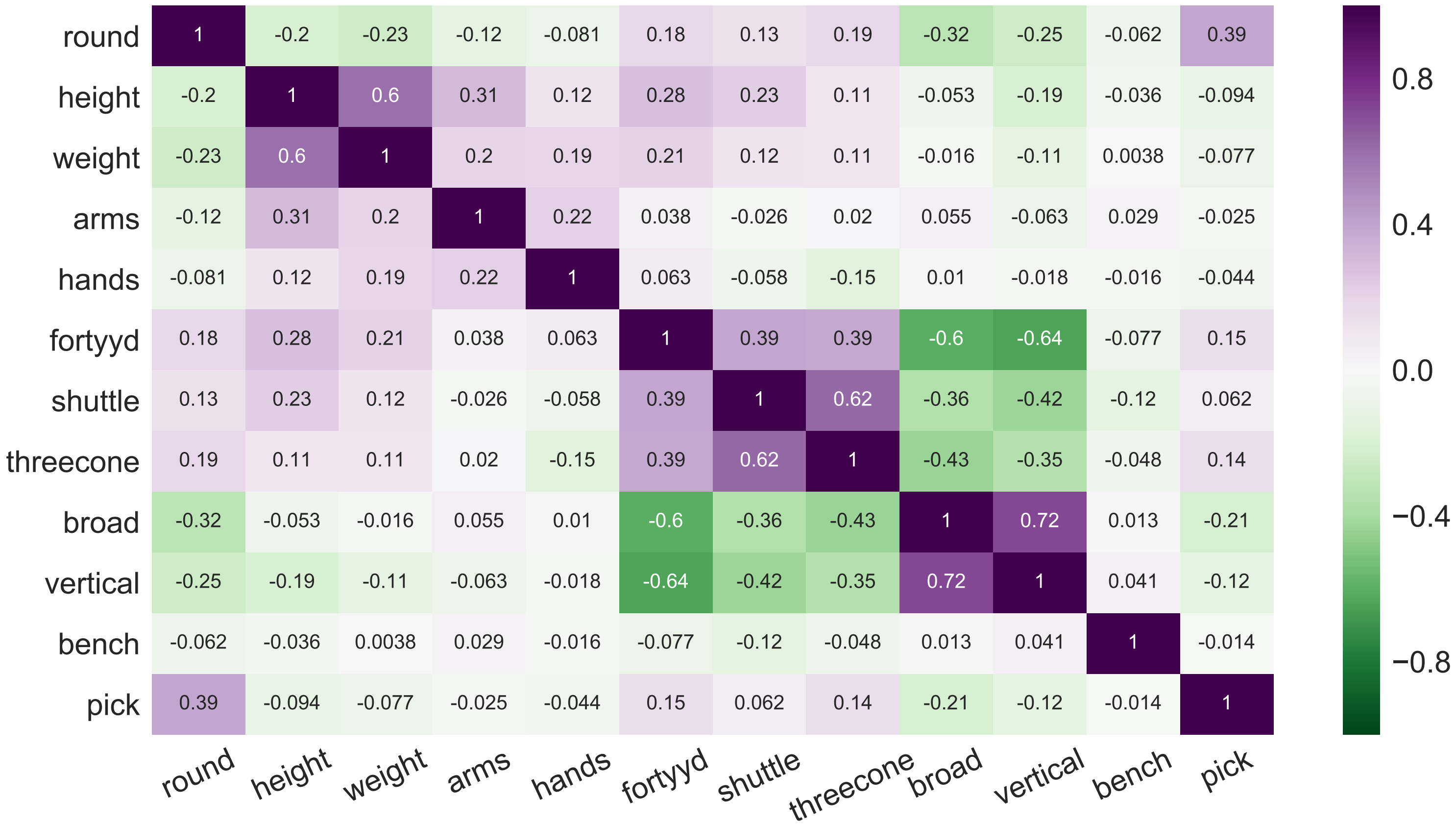

Figure 5: Correlations for drafted quarterbacks.

The correlations between these variables seem to rise when we require the quarterbacks to have been drafted. However, is this a real effect or just a coincidence since we reduced the number of players by 42%? It might be a little of both and perhaps we can explore this effect at a later time. We also must remember another limiting factor in this type of visualization: draft round should probably be a categorical variable. As such, calculating the correlation between draft round and another (numerical) variable is not so easy. Regardless, as long as we are aware of this fact we can still derive some useful meaning.

At the end of the day however, our target variable is the draft round. So it would be good to see correlations in that row (again with the limitations in mind). We can choose a few to look into greater detail by generating a "joint plot" to visualize these bivariate distributions. Let's take closer look at how the broad jump and round data are related for quarterbacks.

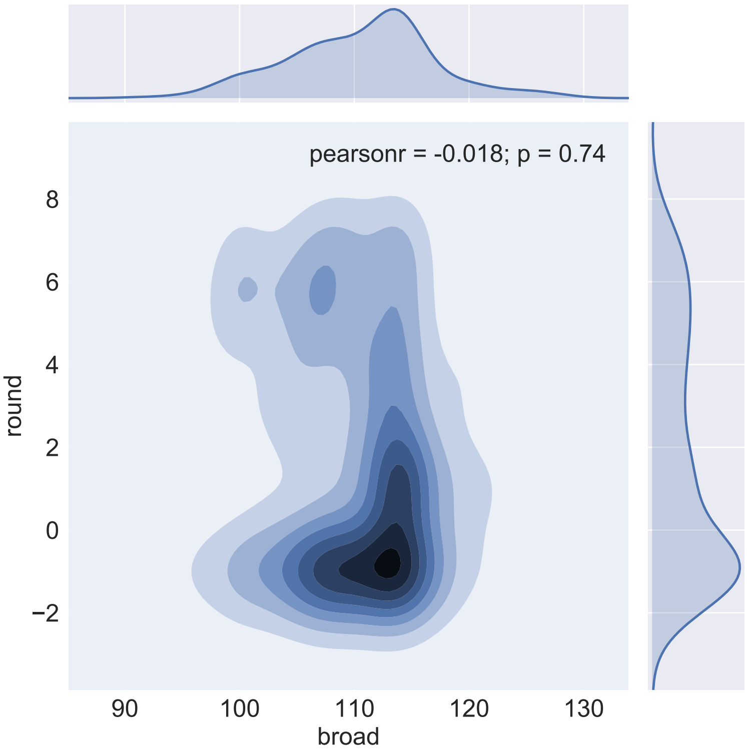

Figure 6: Draft round vs broad jump for all quarterbacks.

A kernel density estimation (kde) is used to help with the visuals of this type of plot. We can see how the concentration of undrafted quarterbacks affects this relationship: the clustering of undrafted quarterbacks has a large effect and a lot of the undrafted quarterbacks have a similar broad jump. A lot of the undrafted quarterbacks have a similar broad jump, with the imputation perhaps playing a large role. Since we want to predict the draft round variable, correlations like these are necessary. However, the draft round is a categorical variable, even if it ranges from -1 to 7 (no zero). This means the usual correlation calculation, the Pearson correlation coefficient, is not truly appropriate to use this case. Instead we can use the Kendall rank correlation coefficient to determine correlation. This test measures the ordinal association and is better suited for our purposes than the Pearson correlation coefficient.

Statistical Inference

As a statistical test, the (Kendall) correlation also gives us a p-value. The null hypothesis in this case is that there is no correlation between the two variables being tested. So a low p-value suggests the calculated correlation is significant. We can assume the usual statistical significance level of 5%. However, in order to really trust the result of this test we can perform a permutation test of the correlation. This permutation test randomly exchanges the draft round labels and re-calculates the correlation. This permuted correlation is calculated many times to develop a histogram (see Fig. 7). From this we can determine a permuted, or estimated, p-value by dividing the number of permuted correlations greater than the observed correlation by the total number of permutations done. This estimated p-value can then be compared to the observed p-value to determine which correlations are significant and which ones are not.

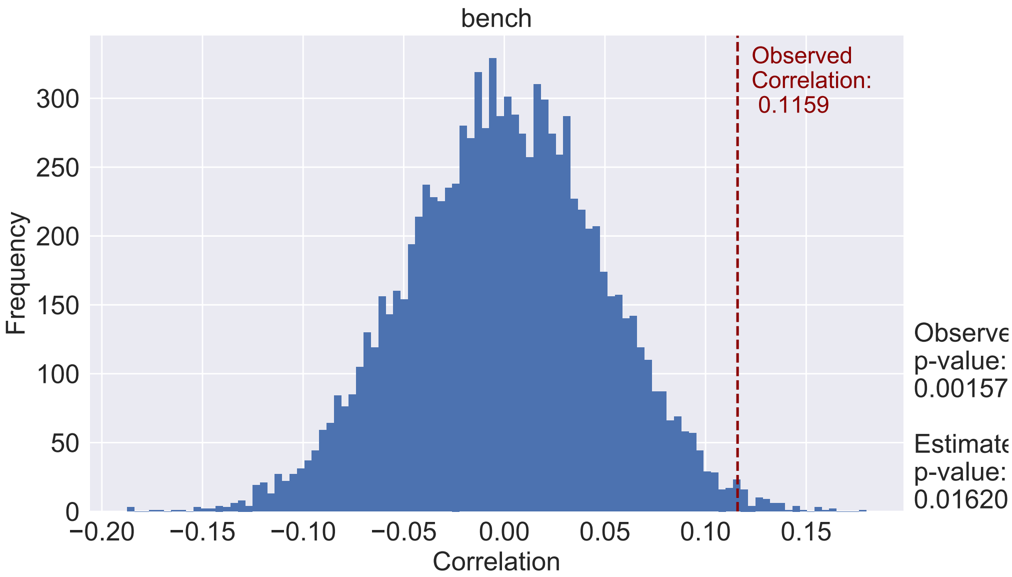

Figure 7: Permuted correlations between draft round and bench press.

After performing this test across the twelve variables of interest, only the height, bench and nflgrade columns show significant correlations to draft round. The bench correlation test is shown in Fig. 7. However, it must be noted that the highest measured correlation is about 0.12, which is quite weak. Regardless, even if they are weak correlations, they still seem to be real, significant ones.

The previous test focused on all of the quarterbacks. What if we separate the quarterbacks who were drafted from those who went undrafted? Are there differences in these types of players? For this test, we can measure the difference in means of each variable of interest. Then we can calculate a z-test and its associated p-value to get the statistical significance. The null hypothesis in this test is that the difference in means between the drafted and undrafted quarterbacks is zero. A small p-value (i.e. $\alpha = 0.05$) implies that we can reject the null hypothesis. When performing this test, we determine that in most cases the p-value is smaller than 5%. There are only three variables with a p-value larger than 0.05: threecone, arms, and hands. Therefore, for these columns, we cannot reject the null hypothesis. However, in all others, we likely can. This means in general, there is a statistical difference between quarterbacks that are drafted and those that are not drafted. An example variable of interest, the shuttle (run) is shown below.

```
   shuttle | z-test statistic :-3.125969
           | z-test p-value   : 0.001772
           | 95% Conf Int     :-0.088894 to -0.020380
-----------|
```

We can expand our focus beyond just quarterbacks and include all players as well. By dividing players into their positiongroups, we can compare how the distribution of a variable of interest is different in any two positiongroups. We can again use a z-test to get the significance in the difference in means of the distributions. For example, we can compare how a defensive lineman's (DL) bench press distribution agrees with a linebacker's (LB) bench press distribution. A z-test and its p-value can help us determine whether there is a statistical difference or not. We can go further and loop through all combinations of positiongroups for each variable of interest. The result is a heatmap of the p-values of the z-tests mapping each of these possibilities. In this type of plot, a low (light) value implies statistically signficant, while a high (dark) value implies no significance. The null hypothesis here is that the difference in means is zero. Remember each variable of interest generates a separate heatmap. The max color scale value is capped at 0.1 to ensure all dark colors land outside the α level we have set (at 5%). This heatmap is also symmetric by construction. An example of such a heatmap is given in Fig. 8.

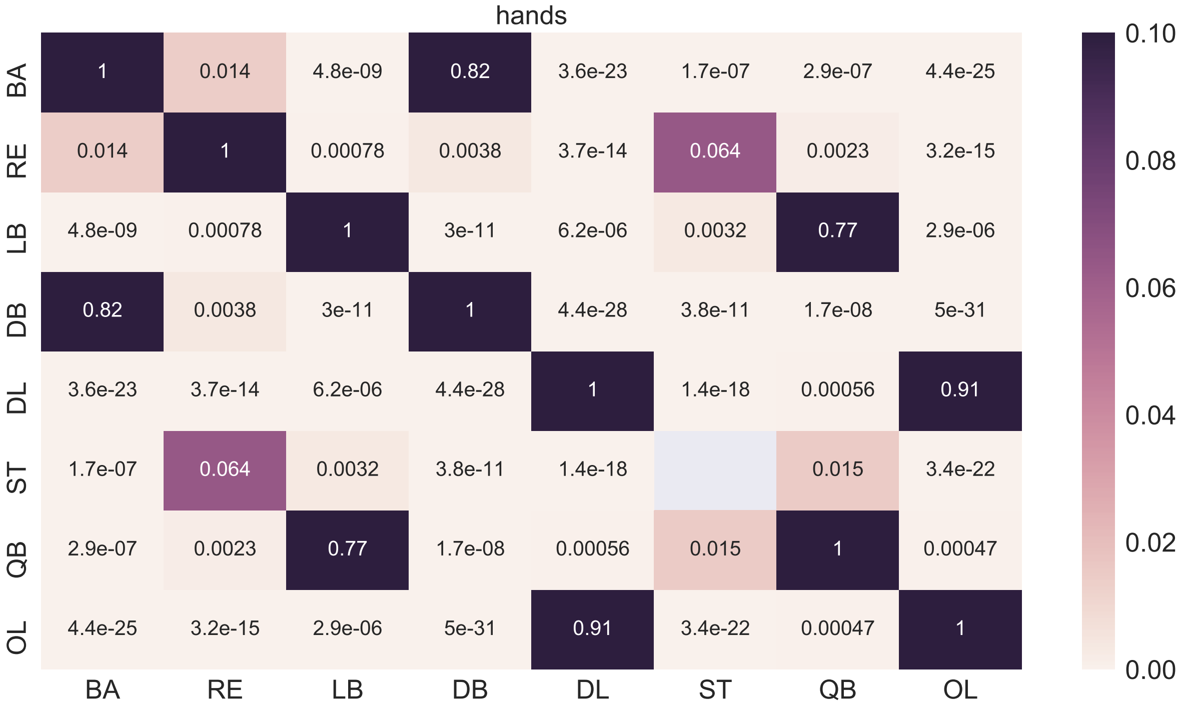

Figure 8: z-test p-values for hands variable for all positiongroups.

From this example, we can see how pairs of positiongroups interact with each other for the hands variable. Offensive linemen and defensive linemen share a very similar mean as it has a p-value of 0.91. However the hands means are very different for offensive linemen and quarterbacks as the p-value is 0.00047, well below our 5% significance threshold.

Conclusion

While some of the correlations are not as strong as we might have hoped, there are certainly indicators that the differences in the variables in the dataset are statistically different. This means there should be ways to leverage that information to make predictions about players and their draft round. The entire project can be found on github here. The code and scripts can be run out of the box as the repository includes all related code, data, and reports, including this document.