**Entertainment Yearly**

**Stat 440: Group 10**

**Introduction**

Our data set features body measurements and the results of athletic tests conducted by National Football League for all players drafted into the NFL from 1999 to 2015 (<http://www.nflsavant.com/dump/combine.csv?year=2015>). The nfl draft features hundreds of college athletes per year and scouts for all teams look for measurable physical attributes as a predictor for success in the nfl. While there have certainly been great players that did not perform well at the combine, many that performed well have found success in the league. We were interested in preparing this data for analysis for the opportunity to see how these measurable traits translated to a player’s success.

**Methods**

With a little under 5,000 observations, there are also twenty-six categorical and numerical variables. Categorical variables include position, year, and college. Numerical variables include heightfeet, heightinches, heightinchestotal, weight, arms, hands, fortyyd, twentyyd, tenyd, twentyss, threecone, vertical, broad, bench, round, pick, pickround, pick total, wonderlic, nflgrade.

There were several issues with the raw data set that needed to be addressed in order to properly read in the information. A small issue was the use of both ‘OC’ and ‘C’ to represent the same position, while the biggest issue related to players who had names with suffixes. In our csv file, the data was input incorrectly because of the additional term and bumped correct values for each variable backwards. Also, there are many players who are missing their colleges. Additionally, each player’s total height needed to be calculated correctly from the feet and inches terms. Lastly, pick needed to match the format of pickround(picktotal).

Since there were only two players who had names with suffixes, we did a manual correction for the entries ‘Mario Edwards Jr.’ and ‘Dante Fowler Jr.’ which fixed their name and all numeric scores for athletic tests. An IF-statement was used to check the total height calculation, replacing any incorrect values, a DO-loop was used to correct values of pick to match the expected concatenation of pickround(picktotal), and finally, any positions listed as ‘OC’ were changed to ‘C’.

After cleaning up the raw data a bit, we noticed several interesting trends and decided to subset the data based on these trends. One subset dealt with observations missing a lot of numeric data, another dealt with players who were missing their college name. We also created several subsets based on shared characteristics such as ‘all running backs’ or ‘all players drafted in BLANK year’ and even physical attributes like ‘players at least 6 feet tall’.

Subsetting data was not the only method of data manipulation we used. We also experimented with conditional output, judging based on nflgrade, and merging data sets (when we added the data set of players drafted in 2015 onto the data set of players drafted in 2014).

**Results:**

**/\*Results section:**

**Charts and tables**

**pertaining to validation and cleaning.**

**Write-up of the results.**

**Point out notable information from the charts and tables.**

**\*/**

Initially, we noticed a great deal of missing/zero values in the dataset. When looking at the frequency tables of each variable we noticed several with a great deal of observations missing. Arms/hands measurements had 76% of observations missing, twentyyd had 97% missing, tenyd had 91% missing, threecone had 79% missing, wonderlic had 95% missing, and nflgrade had 85% missing. The primary reason for the missing values is that some tests were introduced recently and others were discontinued, causing different variables for different players. We ultimately decided not to remove these variables however, so that we were able to subset the observations as described above and still make use of all the variables.