Final Report

I. Summary

This study examines historical NFL player data to determine what factors have historically been overlooked predictors of how good a player will turn out to be relative to their expectations. The variables used were primarily physical attributes that are measured at the NFL draft combine in addition to several categorical variables such as the players position, NFL team that drafted them, and what round they were selected in. To examine the relationship between these factors and the players performance, I tested both Hierarchical models and linear models before ultimately landing on a multiple linear regression model.

II. Introduction:

In this assignment I aggregated various data sources to look for relationships between popular physical attributes used to forecast draft prospects ability and their on field performance relative to their expectations. I quantified the players on field performances by using Pro Football Focus's player rating data, and set their expected performance by what pick they were drafted. I calculated the "expected value" for a player taken at each pick in the draft by taking the player grade for every player selected at that pick in the draft. I used a rolling average for the 10 picks around them to smooth out the variance caused by thin sample size. Then, I subtracted the expected player grade based on draft position from every player's career player grade to generate a value score, with 0 meaning they met expectations of where they were drafted. This value score for each player is what I ultimately use for the dependent variable in my analysis.

In the NFL, there is an annual "Draft Combine" where hundreds of college football players who have submitted their candidacy to the NFL draft are invited to come showcase their abilities. At this combine, they perform a number of drills designed to measure their athleticism and physical attributes. These measurements are believed to have some predictive power on how the player will translate to the NFL.

In addition to the physical attributes measured at the draft combine, I want to look at other variables such as the team that drafted them, the round they were selected in, and the position they play. All three of these factors could potentially have important roles in determining how well a player performs. The team that drafts them has the important job of developing the player, and their ability to develop a given player varies with the coaching, training and support staff at a given organization. The round each player is drafted in seems like it should be a pretty good indicator of how good a player will turn out to be, as so many resources are invested in scouting players to make the best possible selection. However there is much more variation on where good players come from relative to a sport like basketball where the vast majority of stars are taken fairly early in the first round. There are countless examples of elite football players being taken in the 5th or 6th round, and many more that did not even get drafted. Finally, position will be important for using it to group the physical attributes. Different positions prioritize various attributes, so hypothetically grouping by position group might reveal much more than just looking at these traits without prejudice across all players.

III. Data

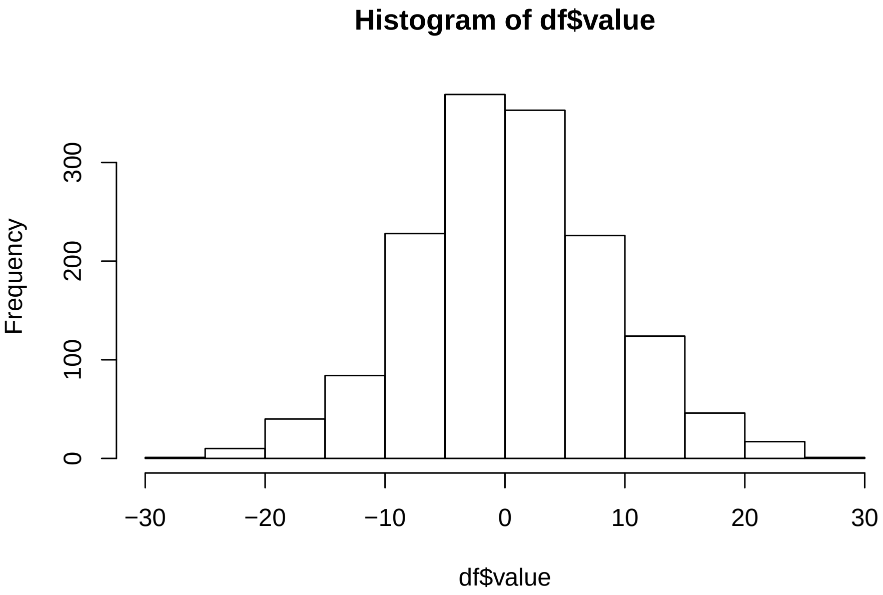
The data I used in this study came primarily from Pro Football Focus and Football Reference. I used Pro Football Focus for the player performance, and Football Reference for draft information and combine measurements spanning the years 2010 to 2019. The players performances were quantified using Pro football Focuses player rating data. This player grade data is very highly regarded and is used by professional gamblers, media organizations, and even NFL teams. A yearly subscription generally costs $200+ dollars but the team at PFF was kind enough to grant me free access for a month. Each player is graded for every snap of every game they play in. The graders are full time employees made up in large part of former NFL scouts, and analysts. Each grade is reviewed by multiple people. PFF adds much more context than traditional stats. Traditional statistics cannot differentiate between a simple screen pass that a receiver runs for 50 yards and a perfectly thrown pass into double coverage that the receiver catches. These both appear as 50 passing yards on a score sheet, but the second throw is exponentially harder to execute from the Quarterbacks perspective. Each player is given a grade of -2 to +2 in 0.5 increments on a given play with 0 generally being the average or “expected” grade. The zero grade is important as most plays feature many players doing their job at a reasonable, or expected, level, so not every player on every play needs to earn a positive or a negative.

Joining these two sources together was an extremely challenging and time consuming process. Profootball Focus data is very guarded and unavailable for extraction and manipulation. I worked with the PFF support staff to get access to the data, however the information I needed was spread over many different worksheets and datasets. Aggregating and organizing all these datasets together presented a unique set of challenges, and a lot of cleaning, name normalization and general work with the Regular Expression package in python. Merging the data from PFF with the information from Football reference was even more difficult. Ten years worth of draft information leads to many duplicate names, mismatched abbreviated names, and inconsistently punctuated names for the ones with hyphens, periods and apostrophes in them.

The data had no missing values for Round, Position, Team, or Player grade, but the draft measurements were a little spottier. Draft prospects will sometimes decline to participate in an event, and as a result somewhere between 10%- 15% of entries had missing rows for a given measurable variable I used. There were a number of players in the dataset who declined to participate in any of these drills. Since these players were missing so many columns, I excluded them from the dataset. For the remaining players that were missing values, I did not want to populate them by simply taking the population averages, because there is too much difference between players for these different drills. For example, I would not expect an offensive lineman to be nearly as fast in the 40yd dash as a defensive back, just like I would not expect a defensive back to be nearly as many bench reps as the average offensive lineman. Similarly, I did not think assigning players values based on players that “look” like them with similar draft rounds, heights and weights would solve this problem either. In my opinion, the best option for populating these values was to take the mean for each column based on the players position.

After cleaning the data to the point it was ready to be used for analysis, I performed a general EDA. A histogram revealed that my response variable player grade has a very even distribution as shown below.

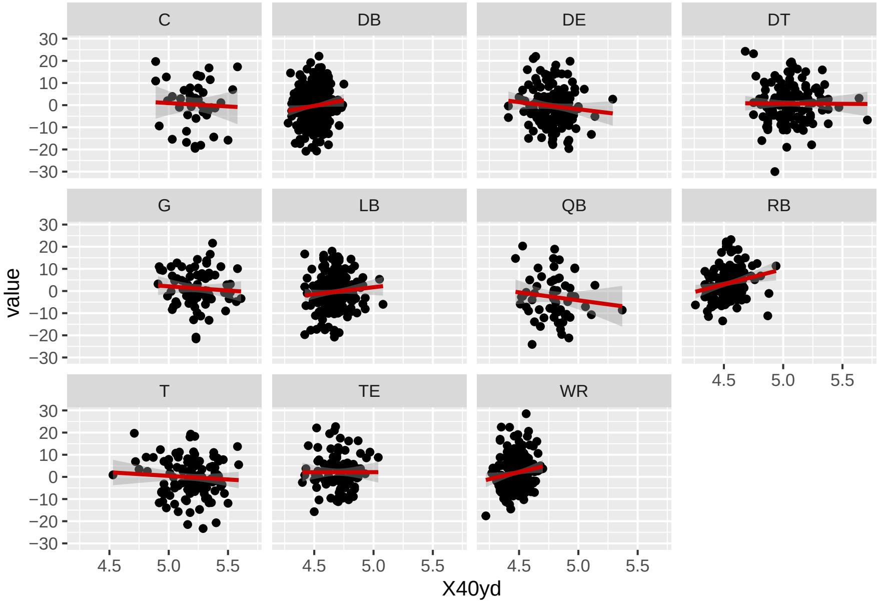
```
hist(df$value)
```



My response variable is “value” and is the delta between a players player grade and their expected player grade based upon draft position. It is a continuous variable that should have an average of roughly 0. From my initial exploration, it did not appear that it would require a transformation. For my independent variables, I factored Round, Team and Position as these three are categorical. Additionally, I mean centered 40yard dash time, bench press reps, height and weight before putting them into a model to help with interpretation, as having a height or weight of 0 does not make much sense, and 40 yard dash times and bench reps of zero will probably lead to some strange intercepts in any interactions they are included in.

One of the most interesting graphs I found in my EDA was the relationship between value and 40yard dash time, binned by position. The best fit lines for each position do not appear to be consistent. Defensive Backs, Wide Receivers, and running backs all have lines with increasing slopes for a slower 40 yard dash time while Defensive End and Quarterback have decreasing slopes with a slower 40 yard dash time. Although it will require further analysis to discover if this is statistically significant, it appears that speed could be overvalued for Defensive Backs, Running Backs and Receivers while it is undervalued for Quarterbacks and Defensive ends.

```
ggplot(df, aes(x=X40yd, y=value)) + geom_point() + facet_wrap(~Position) + geom_smooth(method="lm",col=
```



IV. Model

I attempted quite a few iterations of different linear models, both hierarchical and basic multiple linear regression. I used a baseline model with all variables, and the interaction between position and everyone of my continuous variables. Ultimately, I believe the multiple linear regression model shown below gave me the best results and fewest problems with model assumptions.

This final model has an R-square of .08. This may seem low, but I would expect to be able to be able to explain many strong predictors of generating value in this context with as much investment and brain capital as there already is invested in the NFL draft. I arrived at this final model by comparing R-squares, as well as checking assumptions, running a stepwise AIC test in both directions, and comparing lots of models with ANOVA tests. The model that generated the lowest AIC contained the independent variables of Position, Bench, Broad Jump and 3Cone. When I used ANOVA tests to check for significance in these variables and the ones excluded however, I reached different results. Changing only one interaction or variable at a time from a baseline model to a model without the variable I was testing, I found that the model did not think broad jump, or 3 Cone were significant variables from my AIC generated model. Conversely, it found 40 yard, Height, and the interaction between Position and 40 yard to be significant. An ANOVA test comparing a model to a model without Team found the p-value to be .07. Although it may not be significant at the .05 level, one of the primary questions I wanted to answer is if certain teams have beer better at identifying value at a statistically significant level, so I kept the variable.

$$y_i = \beta_0 + \beta_1 Team + \beta_2 Position + \beta_3 Round + \beta_4 HeightCentered + \beta_5 Round + \beta_6 40yd + \beta_7 Position : 40yd$$

The output of my final model can be seen below:

```r
model4 <- lm(value ~ Position + HtCent + X40ydCent + BenchCent +  Team + Position:X40ydCent, data = df)
pander(summary(model4))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.683	4.17	0.4035	0.6866
PositionDB	0.8592	4.32	0.1989	0.8424
PositionDE	-1.355	4.047	-0.3347	0.7379
PositionDT	-0.8542	4.214	-0.2027	0.8394
PositionG	2.011	5.013	0.4012	0.6884
PositionLB	-1.611	4.06	-0.3967	0.6916
PositionQB	-2.778	4.161	-0.6676	0.5045
PositionRB	4.606	4.276	1.077	0.2816
PositionT	-0.7451	4.433	-0.1681	0.8665
PositionTE	1.468	4.093	0.3586	0.7199
PositionWR	4.99	4.471	1.116	0.2646
HtCent	-0.2466	0.1345	-1.833	0.06702
X40ydCent	-2.975	7.838	-0.3795	0.7044
BenchCent	0.1335	0.05043	2.647	0.008216
TeamBears	-1.694	1.679	-1.009	0.3132
TeamBengals	-1.003	1.569	-0.6394	0.5227
TeamBills	-1.15	1.587	-0.7247	0.4688
TeamBroncos	-0.3368	1.573	-0.2141	0.8305
TeamBrowns	-2.888	1.485	-1.945	0.05199
TeamBuccaneers	-1.868	1.628	-1.148	0.2514
TeamCardinals	-3.493	1.616	-2.161	0.03082
TeamChargers	-0.6018	1.686	-0.357	0.7211
TeamChiefs	0.1915	1.646	0.1164	0.9074
TeamColts	0.8931	1.57	0.569	0.5694
TeamCowboys	1.827	1.604	1.139	0.255
TeamDolphins	-1.696	1.605	-1.057	0.2907
TeamEagles	-0.4644	1.581	-0.2937	0.769
TeamFalcons	-0.2642	1.666	-0.1586	0.874
TeamGiants	-1.118	1.629	-0.6863	0.4926
TeamJaguars	-2.576	1.761	-1.463	0.1438
TeamJets	-1.365	1.603	-0.8519	0.3944
TeamLions	-0.9962	1.607	-0.62	0.5353
TeamPackers	0.1025	1.539	0.06659	0.9469
TeamPanthers	0.9178	1.714	0.5353	0.5925
TeamPatriots	2.646	1.688	1.567	0.1172
TeamRaiders	-1.308	1.561	-0.8383	0.402
TeamRams	-0.3872	1.659	-0.2333	0.8155
TeamRavens	0.7631	1.537	0.4966	0.6196
TeamRedskins	-0.1392	1.586	-0.08775	0.9301
TeamSaints	2.277	1.774	1.283	0.1997
TeamSeahawks	1.04	1.587	0.6555	0.5122
TeamSteelers	0.7799	1.607	0.4852	0.6276
TeamTexans	-1.278	1.578	-0.8097	0.4182
TeamTitans	-0.5597	1.604	-0.3489	0.7272
TeamVikings	1.029	1.598	0.6443	0.5195
PositionDB:X40ydCent	12.57	9.631	1.305	0.1922
PositionDE:X40ydCent	-3.14	9.291	-0.338	0.7354
PositionDT:X40ydCent	1.917	8.788	0.2181	0.8274

	Estimate	Std. Error	t value	Pr(>|t|)
PositionG:X40ydCent	-1.552	9.75	-0.1592	0.8735
PositionLB:X40ydCent	8.101	9.385	0.8633	0.3881
PositionQB:X40ydCent	-3.647	10.02	-0.3638	0.716
PositionRB:X40ydCent	17.81	9.881	1.803	0.07166
PositionT:X40ydCent	2.303	8.826	0.2609	0.7942
PositionTE:X40ydCent	3.34	10.21	0.3272	0.7436
PositionWR:X40ydCent	17.59	10.26	1.714	0.08673

Table 2: Fitting linear model: value ~ Position + HtCent + X40ydCent + BenchCent + Team + Position:X40ydCent The intercept of my model carries the assumption of a player who plays at the Center position for the 49ers with average number of Bench reps, height and 40yard dash time. For this player, my model predicts a value of 1.7.

Observations	Residual Std. Error	R^2	Adjusted R^2
1499	8.032	0.0759	0.04134

Ultimately, only bench press reps, and being drafted by the Arizona Cardinals are significant at the 0.05 level. However there are a handful of variables that are extremely close to having a p-score of .05 or lower.

Bench press has a coefficient of 0.13 and is statistically significant with a p-score of 0.01. In the context of this model, this means that if all other variables are held constant, every additional bench press rep a player can do increases their expected value by 0.13 points. This is an interesting finding because it implies strength is an under-appreciated metric when evaluating players. Although I do not know how NFL teams evaluate players, from the media coverage of the draft combine it seems that a lot of attention is given to a player's athleticism in terms of their high, jumping ability, and 40 yard dash speed. It could be that with all the focus on these flashier traits, how strong a player is can get overlooked and become underappreciated.

Being drafted by the Arizona Cardinals has a coefficient of -3.49 with a statistically significant p-score. This means that holding all other things equal, being drafted by the Cardinals reduces your expected value by 3.5 points. This carries the implication that the Cardinals are especially poor at drafting and developing players. Although the Cardinals have not been a terrible franchise from 2010 to 2019, they have certainly been below average and much of their success has come as a result of players they acquired as a result of free agency rather than the draft. It is worth noting that the Cleveland Browns have a coefficient of -2.88 and just missed out on statistical significance with a pscore of 0.052.

Although they were not significant at the .05 level, the interactions between 40 yard dash and running back and wide receiver just missed the cut. They have coefficients of 17.81 and 17.59 respectively meaning that a 1 second increase in a running backs 40 yard dash speed would result in an expected increase in player value of 17.81, and an increase in player value of 17.59 for receivers. Although this initially appears quite large, a one second increase in 40 yard dash speed at these positions makes a massive difference. The vast majority fall in a small range of values from 4.5 to 4.8 seconds. Speed is also a trait that logically seems like it should have a negative relationship with how well a player performs, but as the response variable is value rather than player grade, I think this is an interesting finding. Speed is one of the flashiest traits at these positions, and every single year a player's draft stock increases dramatically when they run a fantastic 40 yard dash. However, speed is just one aspect that goes into a players ability, and it does not surprise me that it receives far too much focus when a player is being evaluated.

The 95% confidence intervals for the Cardinals and Bench reps can be seen below. The Cardinals ranges from -6.66 to -0.33. This means that we are 95% confident the true effect of being drafted by the cardinals has on player value falls between -6.66 to -0.33. The Bench coefficient spans from 0.03 to 0.23, meaning that we are 95% confident the true effect of every additional bench rep is an increase in value of 0.03 to 0.23

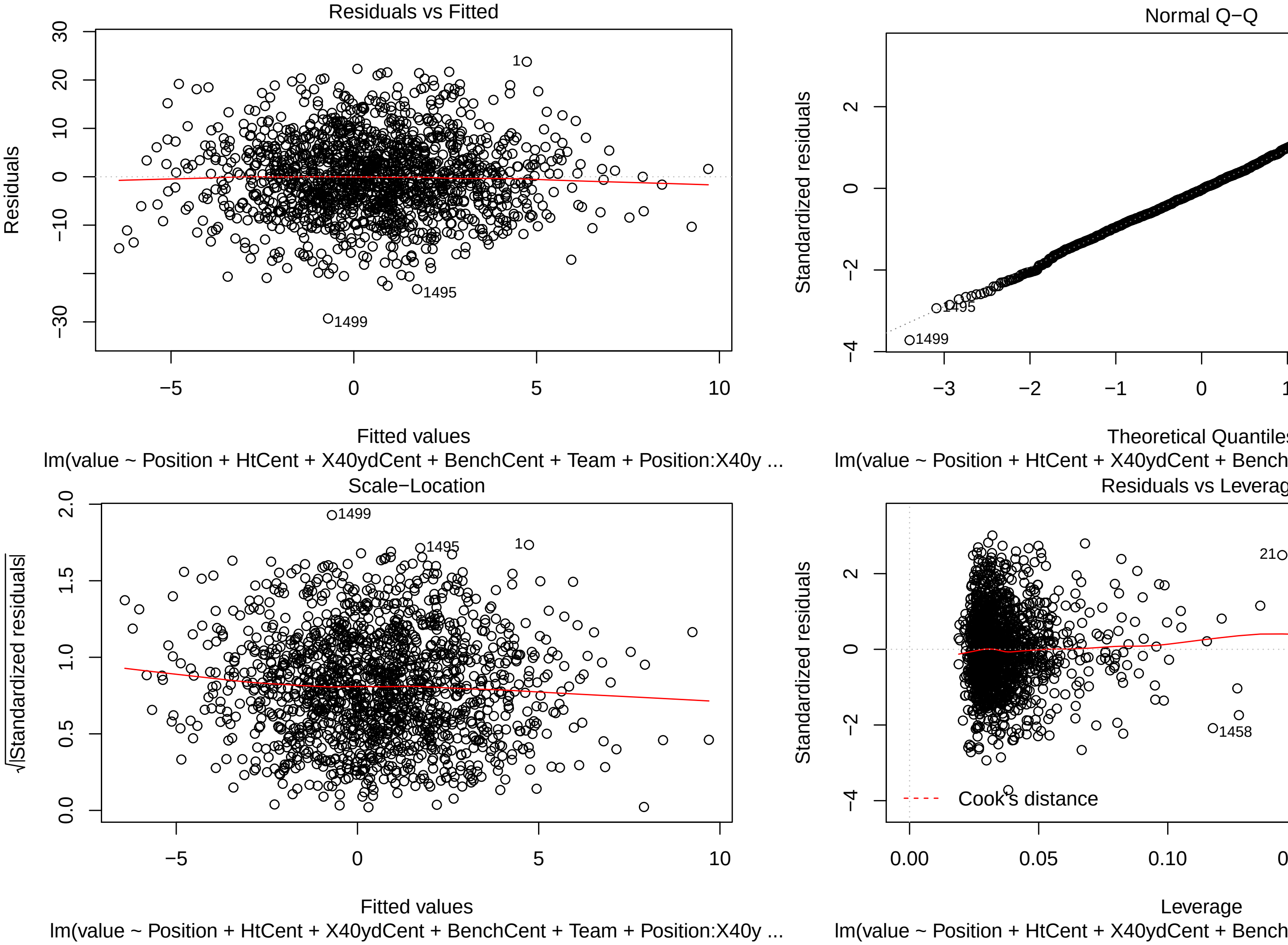
```
pander(confint.default(model4, c('TeamCardinals', 'BenchCent')))
```

	2.5 %	97.5 %
TeamCardinals	-6.66	-0.3257
BenchCent	0.03464	0.2323

```
#vif(model4)
```

There are no violations of linearity, independence, normality, or equal variance in my model. The residuals vs fitted plot is randomly distributed, almost all the points on the Normal Q-Q lie exactly on the 45 degree line, and the standardized residual plot is random as well. From the Residuals vs Leverage graph there appear to be no high leverage points with no points falling outside of the cook's distance line. The VIF scores are all below 5 except for the positions, and 40yd dash times which are all extremely high. This is expected as the interaction term of Position and 40yd will influnce this.

```
(plot(model4))
```



```
## NULL
```

V. Conclusion

This study was attempting to identify what measurables go into finding value in the draft, and what teams are significantly better or worse at identifying and developing talent. Ultimately, my model identified answers to both of these questions. The Arizona Cardinals are worse than the rest of the NFL teams at finding and drafting talent at a statistically significant level, and the Browns are not far off. Strength (as measured by bench press reps) is an undervalued attribute in the draft that appears to go somewhat overlooked by teams. Conversely though it is not statistically significant at a 0.05 level, speed (as measured by the 40 yard dash) tends to be over valued when evaluating running backs and wide receivers. These findings could potentially be of use to NFL teams when evaluating players.

There are limitations to this study however. The findings in this study are based on player grade data, and although this data is highly regarded, it is ultimately the aggregated opinions of a few people, and are not a perfect representation of how good a player is. While more data is always preferable, this is especially true for an analysis like this that includes a variable like team that has 32 categories. This spreads the data and results even further. Finally, there is far more that goes into evaluating a player before the draft than the variables in this study. Their college performance, and personality, and intangibles all play a role in how teams perceive and select players. My model with an R-squared of .08 reflects this fact that it does not come close to capturing the full story of player evaluation.