Lauren Chapman
Kennedy Kiemsteadt
Michael Opiela

Moneyball: NFL

Background and Problem

From Bill James to Lebron James, moneyballing sports has become the new norm. Is the NFL its most challenging conquest yet? Can you "moneyball" football? Our final research report is centered about the phenomena, "Moneyball". Moneyball can be defined as a phenomenon within sports – particularly in baseball – that refers to the use of advanced statistical analysis and data-driven decision making to assemble a competitive team, even with limited resources. In 2003, Michael Lewis' book, "Moneyball: The Art of Winning an Unfair Game" has since popularized the term. Lewis' book follows the story of the Oakland Athletics Major League Baseball team that was led by General Manager Billy Beane. Beane used in-depth statistical analyses to understand how to identify players who were "undervalued" and in turn build a successful team, despite having one of the lowest finances within the MLB. By re-evaluating their strategy in this way, the 2002 Oakland Athletics with "approximately $44 million in salary, were competitive with larger market teams such as the New York Yankees, who spent over $125 million in payroll that season" (Wikimedia, 2023). This approach ultimately brought the A's to the playoffs in 2002 and 2003. Building on the A's original success, other teams in Major League Baseball followed suit, with the Boston Red Sox winning their next World Series soon after, breaking the infamous "Curse of the Bambino", based on these principles. "Moneyball" was eventually replicated in other sports, most notably next being the NBA.

In essence, Moneyball potentially enables teams and organizations to make more educated and informed decisions regarding team-building and player management. Moneyballing has since been utilized across various sports and industries alike as an approach to make data-driven decisions and maximize opportunities and growth. By taking this approach into account, we wanted to determine if Moneyball could be applied to football, specifically in the NFL. In lieu of the 2023 NFL Draft, our team was curious to know if we could leverage our analytical and data-driven skill sets to uncover if moneyball in the league is possible, if it should be done, and if a version of it has already been done.

Our team obtained data from Kaggle, an online platform for data science and open-source datasets. The dataset we collected, “NFL Draft 1970-2021”, was used to perform various in-depth types of analyses such as linear regression models, frequencies, distributions, and new dataframes. This dataset was scraped from pro-football-reference.com. We then created a data frame with important variables that we wanted to analyze, and only included data from 2013-2018 NFL Drafts. With this dataset, we aimed to gain insights and make informed recommendations based on our findings. Below are some questions that guided our framework of our analyses:

Is Moneyball in the NFL possible?

We are investigating this overarching question by using metrics that currently exist within the NFL. Because Moneyball hasn't been implemented entirely across all professional sports and organizations, we wanted to delve into the realm of football and how the approach could be utilized. Since the 2023 NFL draft is currently happening right now, we wanted to use prior statistics to determine possible Moneyball strategies. By doing this, we could frame our hypothesis as our draft strategy. Everyone who has been in charge of an NFL team knows that the higher the draft round, the better the chance at getting a high value player with a given pick. Although the draft looms as largely still a guessing game, there is a lot of information available to make data-driven decisions.

2014 film *Draft Day* follows the story of an “opportunity to rebuild a team when General Manager Sonny Weaver trades for the number one pick” (IMDb). The story revolves around researching the statistics and personality traits of the expected number 1 overall prospect, a quarterback. Although the predetermined first pick had every hard skill you would want in the draft, the player ultimately ended up not being drafted as #1. This was due to the fact that they figured out he does not have the redeeming qualities as a leader and good teammate.

This example, though fiction, holds true in the NFL to this day. An NFL player could have amazing stats and measurable traits through the data set that NFL talent evaluators have, which is why events like the NFL Draft Combine and College Pro Days are made available, because it allows scouts to see their prospects “test” in person to verify their measurables and add more data to the dataset of game film that they already have to evaluate players off of. Though this is more good information, players like Jamarcus Russell, and more recently small groups of players like highly drafted quarterbacks such as Jameis Winston and Marcus Mariota can be amazing in college (both previously playing in national championship games, great stats and game film) and measure well or at least adequate to scouts expectations (they were drafted numbers 1 and 2 that year) but between them, never really developed into the "stars" that they were expected to be. You don't take a quarterback so high in the draft in the NFL without expecting that to be your franchise cornerstone for the next 10-15 years, so anything short of that is seen as a disappointment at the least. But because that type of disappointment is so common due to the flawed nature of current evaluation processes by scouts, we wanted to explore further and see what we could do with the available data. Based on what we see prior to our analysis, Moneyball in the NFL could very well be possible, but not in this current system. This current version of evaluation is not as successful as it needs to be, and can be categorized as deeply flawed entirely if our hypothesis turns out to be false.

Should the Moneyball approach be done in the NFL?

The process of drafting players can often be flawed, and there are many factors beyond statistics that can contribute to a players success. Since we are looking at statistics, we can do the best we can in order to predict the success of a player based on career average over time and by round pick, as well as games played. Using this data, we can help prove that scouts or leaders in the industry can still be wrong as to how successful a player will be, simply based on stats. It is important for these leaders to ask the right questions and transition from the older, more outdated way of drafting players into a more modern and realistic method. This data also proves that some hypotheses of success are correct, and that good player stats often leads to very successful players, but where there are anomalies in the data where players with great stats do not end up doing as well as hypothesized, we can contribute that to possible life changes, how they contribute to the culture and family of the team they are on, attitude, or flaws in scouting that may not have been brought up in the selection process. Therefore, a more overarching analysis of a player should be done, in order to help the player succeed and to help the program they enter thrive.

Has a version of Moneyball in the NFL already been done?

Not as it was implemented in other sports, football has so many variables and differences by positions, by how each team evaluates their prospects/current players, and what they're looking for to fill a need on their team. Also, just because each team has a framework doesn't mean it's a "successful" one, as gm's and coaches get fired all the time, effectively re-setting their drafting and player acquisition strategies. With the 2023 NFL draft occurring right now, we decided to test the statistical effectiveness of the most common nfl draft strategy in nfl history, which is built on the assumption that teams are built around high draft picks, and the higher round your pick is in, the more likely you are to be adding a future pro bowler, all-pro player, or NFL superstar to your team to improve its performance in the near future.for this analysis, that success is measure by the best existing metric available, which is carAV (career approximate value) of a given player. We'll dig into that more in the next section.

What variables from the dataset are best to be used?

Within our dataset, there are numerous tools that could make a good determinant of how Moneyball could be utilized within the NFL. Variables such as carAV (career approximate value) and drAV(drafted team approximate value) are examples of such. We infer that it'll be pretty similar to carAV because a lot of the players who make it to free agency following their first contracts haven't yet had enough time to shift their value away from being majority attributed to their originally drafted team. All_pro, the best of the best players, chosen because any player this young with a notable amount of all_pros will have a good chance to make it to the nfl hall of fame. Pro bowls were measured more as a less rare stat to separate some of the "stars" from your average or even above average players. Though the fan vote component biases this stat towards larger markets and more popular players and teams, it's useful to compare to all_pro to get a more clear picture of the NFL landscape. Games played we used to determine how many players actually contributed at all to their teams, as at times later round picks see the field more as substitutes, and more games played by round also implies that each prospect in a round actually contributed, so the distribution of talent is more consistent and even across a round with more games played, since you can't play more than 16 in a year previous to this past 2022 nfl season, when it was lengthened to 17 games.

How can the variables be improved?

The variables can be improved by finding data that provides more granularity. This would allow us to look at more of the specifics such as game by game and play by play data – similar to the plus/minus and win shares statistics within the NBA. Moneyball in sports can be deemed important due to the fact that it places a high emphasis on the use of data analytics to identify undervalued players who can potentially perform successfully and help the team/organization win games. This approach in turn has completely transformed the way teams evaluate and acquire players, as it focuses on the objective measures of player performance, rather than subjective assessments or popular opinions. By relying on statistical analysis, teams can identify players who possibly were overlooked and/or undervalued by other teams, allowing them to acquire those players at a lower cost. By doing this, teams can potentially create a competitive advantage in an already competitive field. As previously mentioned, Moneyball has proven successful in other sports such as baseball and basketball, where teams have implemented similar data-driven approaches to build winning teams.

We believed that within the MLB and then NBA, Moneyball strategies relied on a wide range of data and quantitative analyses to identify undervalued players and ultimately build successful teams. They created or re-utilized existing stats that they valued more than others as main metrics to determine a player's worth relative to their team's success and what that looked like to them. For the A's, it was on base percentage and slugging percentage, which they used to increase runs scored, which they saw as converting to wins as long as they passed their calculated threshold of runs needed to win, which they converted to wins. Basically, they used a myriad of underappreciated stats to build their own indexed stats to convert to wins. In our analysis, we're using the pre-established stat (refer to link explaining carAV) Career Average Value, which takes into account multiple commonly used stats to build it into one of the few "universal" stats in American Football that can be used to compare player value across different positions groups which largely don't have important stats that can be compared to each other due to the highly specialized nature of the player position groups of American Football. We also included some other commonly used stats for additional explanation and granularity, as needed. Seen below are some of the variables that were seen in the dataset and that we chose, along with the NFL Drafts from 2013-2018 in our created data frame, which we leveraged within our analysis.

1. Weighted Career Approximate Value (carAV)
    a. carAV, known as the "weighted career approximate value" determines the "balancing peak production against raw career totals" (Approximate Value). For each given player, carAV contains the weighted sum, of seasonal approximate value of:
        i. 100% of the player's best season, plus 95% of his 2nd-best season, plus 90% of his 3rd-best season, plus 85% of his 4th-best season, and so on…" (Approximate Value).
2. Drafted Team Approximate Value (drAV)
    a. Very similar to carAV, except this stat explains how much value the player brought only to the team that originally drafted them.
3. Rounds (round)
    a. Rounds 1-7 a player was drafted in

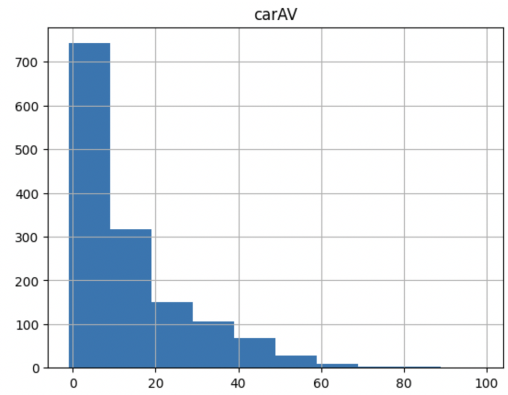
4. All-Pro Teams Made (all_pro)
    a. Number of All-pro teams a player made
5. Pro Bowls Attended (pro_bowl)
    a. Number of Pro Bowls a player attended
6. Games Played (games)
    a. Number of games a player participated in
7. Age (age)
    a. Age of player

Because of the simplicity of the data, more complicated statistical analyses weren’t needed, and considering the different stats with different levels of weight and importance by positions and team, comparing more statistical variables for each player wasn’t viable or possible. It would've been like comparing apples to oranges to potatoes, so we stuck with the established universal metric, and chose a couple other existing but less “perfect” ones to compare.

The use of Moneyball strategies can provide valuable insights and recommendations to the NFL for business strategies and policies to further optimize performance and profitability of teams. Because Moneyball is centered around data-driven decision making, it can be applied to various business decisions such as pricing, marketing, and operations. The NFL can use the data and its findings to optimize their revenue streams and ultimately increase profitability. In addition, Moneyball strategies can help NFL teams identify those undervalued players to make smarter decisions when it comes to talent acquisition. By using data and analytics to evaluate player performance, teams can find players who may have been overlooked by other teams and acquire them at a lower cost, creating an overall competitive advantage. Overall, Moneyball strategies can provide valuable insights and recommendations to the NFL for optimizing performance, improving profitability, and making data-driven decisions that can lead to success both on and off the field.

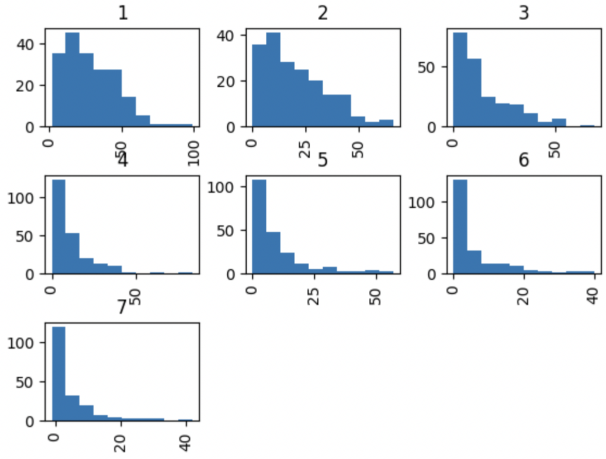
Data Summary and Exploratory Analysis & Data Analyses, Key Findings and Conclusions

Overall Weighted Career Approximate Value (2013-2018)



Across the range of the entire dataset we looked at, we see a heavy right skew which tells us that over the course of these 6 years of draft data, the vast majority of these players have little to no significant value added to the team value. Roughly half of the players whose career approximate value we looked at are from ranges 0-10, which is marginal additional approximate value. This also means that half of these players have added significant value to the teams they have played on throughout the same time period. This tells us that the players that we examined within these years have had a significantly positive impact on the NFL quality of play as a whole. With that being said, these players are worth analyzing and we can dive deeper into our analysis to see where this value lies more specifically.

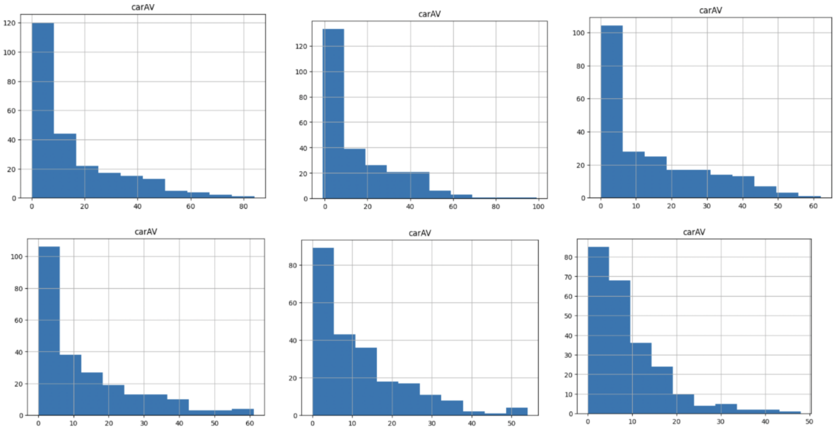
Overall Weighted Career Approximate Value by Rounds (2013-2018)



Within these new sets of graphs, they show a strong right skew around the .0. We can see that the higher the round, the less we see the majority of the picks clumped around 0 approximate added value.

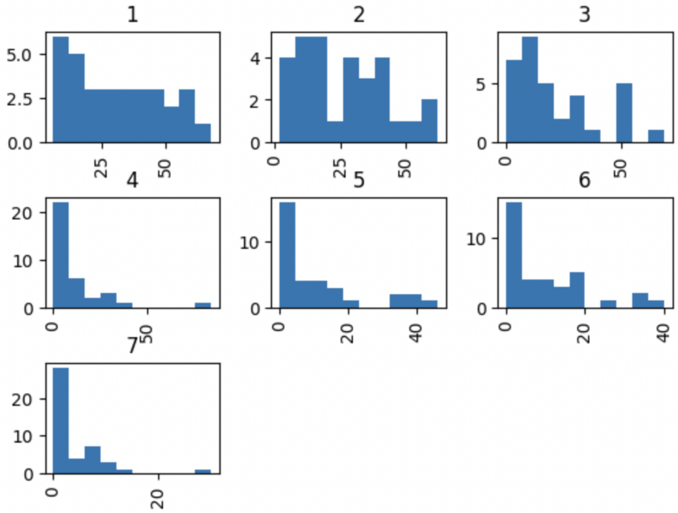
This can tell us based on simply looking at the graphs, the distributions indicate that the higher round you get the higher career approximate value players you'll be adding to your team.

Weighted Career Approximate Value Across Each Year (2013-2018)



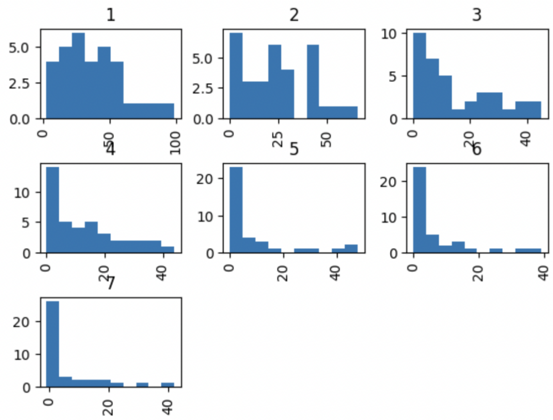
As seen is the weighted career approximate from 2013-2018 reading from left to right on each row. The distributions follow a similar pattern throughout the years and are all skewed to the right at the .0.

Weighted Career Approximate Value by Round in 2013



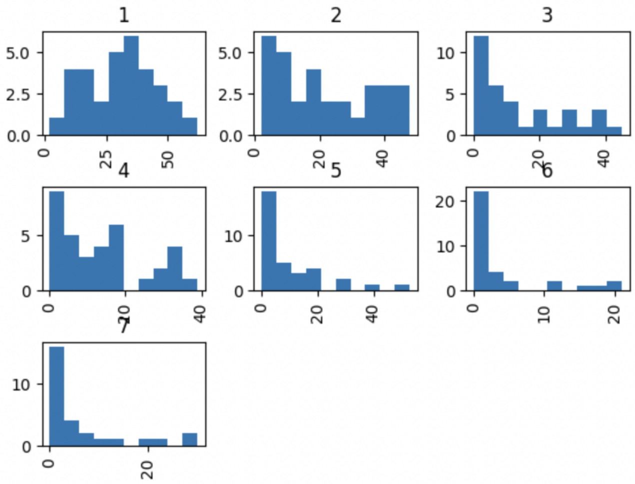
Each round largely shows what we expected – a continuation of the right skew we have seen before. We are expecting a similar trend across the 2013-2018 years in consideration of this variable. But, a noticeable difference we see is in the 4th round we have one outlier that looks like they either had a long career or made multiple pro bowls and all pro teams due to their high value. As you look at each round, the overall trend shows lower career approximate value from the first round all the way down to the seventh.

Weighted Approximate Value by Round in 2014



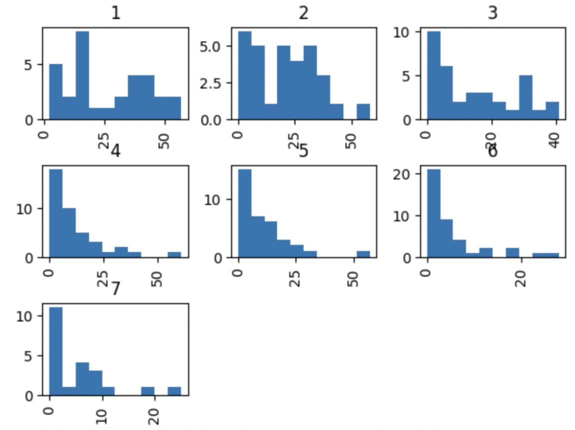
We can see that the first round was an extremely top heavy draft. We have contributors that had almost all the way up to 100 carAV which is significantly higher than what we saw in 2013. We see a rough continuation across rounds of the mentioned right skew trend within the data. The fit of the trend is a lot more rough within the first two rounds. We see a more even distribution across the higher to lower end contributors. With that, we see a much more top heavy but also balanced draft in terms of the first and second rounds.

Weighted Career Approximate Value by Round in 2015



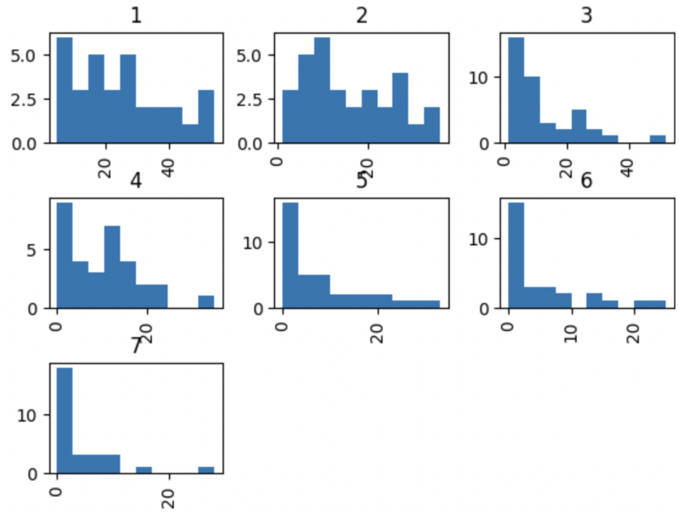
In round 1, we see our first normal distribution. Within this distribution, we see a peak at approximately 30 of carAV for players taken in the first round. This will be interesting to watch to see if the distribution stays close to normal or starts becoming right skewed. In round 2, we see a bimodal distribution meaning that there was a high frequency of both non contributors and higher contributors in this round in 2015. In rounds 3, 5, and 6 there is a return to the expected right skew. Round 4 also appears to be slightly bimodal.

Weighted Career Approximate Value by Round in 2016



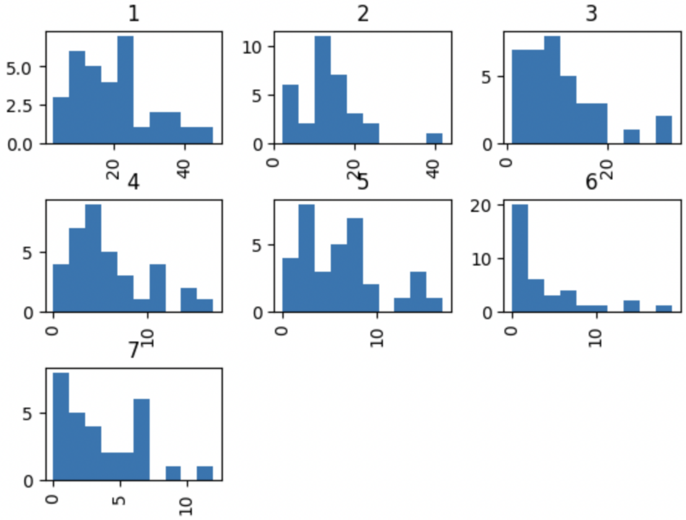
Fortunately, we are back to the right skew trend within the rounds of the 2016 draft. We also see single outliers in rounds 4 and 5 in which both outliers have a carAV of just over 50 which we have rarely seen across this analysis thus far. Additionally, round 2 is evenly distributed with a similar frequency of lower and medium contributors which tapers off towards higher contributors. Though it still follows our expected trend, it is a very rough fit and as players from this round and year continue to develop, we could see a normal or even left skewed distribution as they accumulate carAV in the future.

Weighted Career Approximate Value by Round in 2017



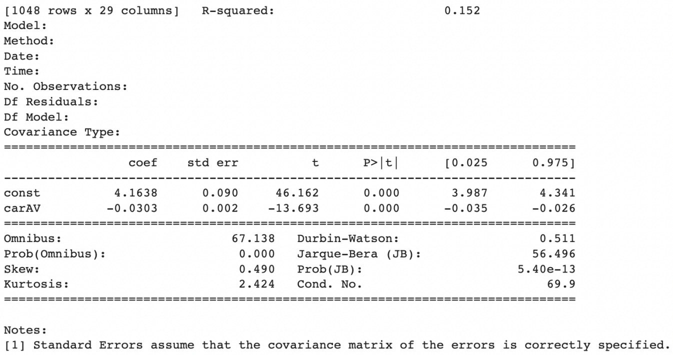
We see yet another continuation of the expected trend of a right skew. However, a notable difference in the 2017 draft versus the previous draft years we have analyzed so far is that many more of these prospects ended up with a lower carAV thus far in their careers. This could be due to the fact that these players were only drafted five years ago and havent come to their prime/full potential. We are starting to see the overall carAV drop off as we analyze players who have had NFL experience but have not fully come into their prime. Additionally, round 2 seems to be approaching a normal distribution already – another specific round top watch in the future as the players picked in that round continue to develop.

Weighted Career Approximate Value by Round in 2018



In 2018, we see a normal distribution for the first round. Because of that, we would want to watch this round moving forward. There are very few players within the first round that fall into the non contributor bar of this histogram, so as we continue our analysis we might want to look into this round more specifically. Throughout the rest of the rounds in this draft, we have an extremely rough fit to our hypothesized trend. This could be because similarly to 2017, there is still a lot of time left for these players to grow and show their value. We see enough frequencies of players towards the higher end of value that there is hope for this draft class. We will likely continue to see players from this draft class over the next few years to gain a complete picture of this NFL draft class.

Linear Regression #1

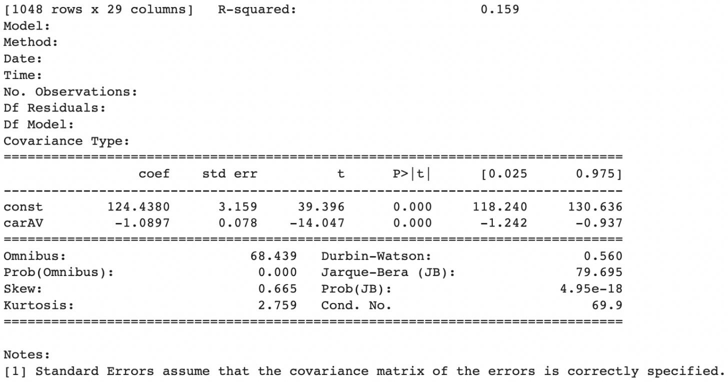


```
[1048 rows x 29 columns]   R-squared:                       0.152
Model:
Method:
Date:
Time:
No. Observations:
Df Residuals:
Df Model:
Covariance Type:
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.1638      0.090     46.162      0.000       3.987       4.341
carAV         -0.0303      0.002    -13.693      0.000      -0.035      -0.026
==============================================================================
Omnibus:                       67.138   Durbin-Watson:                   0.511
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               56.496
Skew:                           0.490   Prob(JB):                     5.40e-13
Kurtosis:                       2.424   Cond. No.                         69.9
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

The r-squared in our simple linear regression tells us that only 15% of the variation in Career Approximate Value is explained by the Round that each player was drafted in. This tells us that purely basing off of each round, you're still dealing with significant uncertainty in terms of if you're getting a significantly positively contributing player to your team. From this, we know that it's not advisable to base your value of a player solely on the round they were picked in.

Linear Regression #2

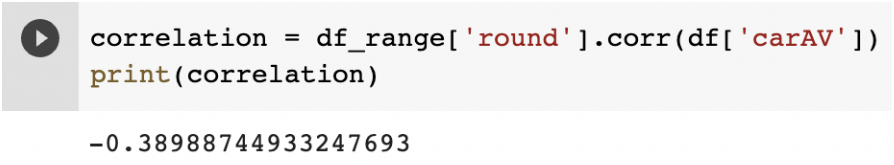


```
[1048 rows x 29 columns]   R-squared:                       0.159
Model:
Method:
Date:
Time:
No. Observations:
Df Residuals:
Df Model:
Covariance Type:
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        124.4380      3.159     39.396      0.000     118.240     130.636
carAV         -1.0897      0.078    -14.047      0.000      -1.242      -0.937
==============================================================================
Omnibus:                       68.439   Durbin-Watson:                   0.560
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               79.695
Skew:                           0.665   Prob(JB):                     4.95e-18
Kurtosis:                       2.759   Cond. No.                         69.9
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Based on the results of our first regression, we thought that basing the analysis on what pick the player was picked with would change our results to something more insightful, but this was not the case. Since the results of this regression mirrored the first, we have the same conclusion, except also we know that basing a player's value on draft position whatsoever without any other variables means you should definitely lose your job as an NFL general manager.

Correlation Results (round x carAV)



```
correlation = df_range['round'].corr(df['carAV'])
print(correlation)
```

```
-0.38988744933247693
```

As we hypothesized, the higher the round (highest being 1) the higher the career average of the player will be. Therefore, when the rounds lower, the correlation decreases because lower rounds will have players with lower career averages

We determined that the most successful way to conduct our analyses would be to focus on a given time period within our data set that spans over 50 years of the NFL Draft. We focused on the years 2013-2018 data to focus on players who are currently playing in the NFL right now. This range gives us roughly five years of data with a statistically significant amount of games played. We chose 2018 as an end year because the 2018 Draft Class has the most complete data, and the most recent class with a complete data set from their full rookie contract. The rookie contract is important because length is negotiated in a way to where that's how much time players and teams agree that it generally takes a player to acclimate to the NFL and grow into an NFL player. By this time, it is generally known that most players see their value in the open market in an NFL free agency. These players can either take an extension or get signed by another team for their second contract. Additionally, this specific time period and range accounts for the retirement age ranges. In this time, players have begun to retire at larger numbers or are near the retirement age.

Most of our charts showed a heavy right skew, which tells us that overall, the higher the career approximate average, the more likely a player was to get picked in the first or second round. In addition, when looking at career approximate average over time, it was interesting to see that over the couple years of data we looked at, the career approximate averages of the players added to teams decreased, which may mean that the players they are choosing are adding less value over the span of 2013-2018 on pro teams. This leads us to believe that the scouts were looking differently at talent, or certain statistics were not being evaluated as heavily over the years. Our regression allowed us to be able to recommend that scouts or draft leaders should not base the value of the player based just on what round they were picked in. There were certain anomalies in the data which were listed in our key insights, which allowed us to prove that some years have prospects that end up with lower career averages than expected based on round pick, and that rounds 3, 4, and 5 ended up with much more skewed data, most likely based on player performance fluctuating or playing unexpectedly.

Strategy or Policy Recommendations, Limitations, and Future Research

After analyzing the data, we found that generally, the players with higher stats in the pros were drafted in higher rounds, and the trends do add up, but there are also anomalies. These anomalies prove that there are some players drafted that cannot be meaningfully predicted given the data and processes that currently are available, and with effort and time put into them or being presented with a different team culture than they were previously in, they may flourish more in the pros than originally predicted. When looking into future research, it would be interesting to see how much of a factor the amount of money or quality of fanbase a player will bring to a team plays in the drafting process. Each team wants great players, but the program also wants money. Therefore, if a player has a huge college fanbase, they might be hypothesized to fill more seats in the stadium with new fans that previously were not fans of the pro team.

In addition, we did not have previous data on players before the draft, which could have helped us see where they ranged statistically in college, and where they are statistically in the pros. Using this, it would be interesting to compare success in college to success in the pros, and highlight some potential anomalies that might contribute to lack of overall leadership, work ethic, or attitude from a player that were not seen by scouts. As previously stated, we were given so many variables that we had to just use the variables that were key to our task at hand, but it would be interesting to look at defensive and offensive statistics and success of those players in comparison to one another. We could analyze the success of specific positions on the field to determine what positions end up being most successful in the pros, and how certain positions can be more prone to injury or over time end up spending shorter amount of time in the pros while in other positions, a player may end up staying for a super long time. If we also had salary data, we would be able to analyze salary and stats based on team, and look at players across teams with similar statistics and see how much they are paid based on the team they play for. If we were to do this, popularity of the team, program overall funds, and season success would be factors that would contribute to salary.

Some limitations that occurred included having missing data for some variables, which luckily we were not using for the specific question we were trying to answer. We also were aware that some data was not recorded before 1994, such as defensive statistics, and also that COVID-19 may have had an effect on some players' statistics after 2020. Therefore, we focused mainly on 2013-2018 data, to give us the most accurate statistics to use and analyze trends. In addition, the process of scouting and drafting players might have changed or adjusted over time, and player skill also changes over time, which could have had effects on the data that were undetectable in the analysis process. It is also important to note that the data includes some players that had not reached their full potential in the NFL or had just begun their rookie season and might not have had much playing time. As stated before, we would recommend a more thorough analysis of each player including team and coaches feedback, in order to take a more holistic approach to drafting. Having a fair, open conversation between scouts and decision makers is crucial to the quality of players in the draft, as well as the treatment of the players as well.

Appendices

Approximate Value. Pro Football Reference. (n.d.). Retrieved April 29, 2023, from https://www.pro-football-reference.com/about/approximate_value.htm

Cviaxmiwnptr. (2021, May 13). *NFL draft 1970-2021*. Kaggle. Retrieved April 29, 2023, from https://www.kaggle.com/datasets/cviaxmiwnptr/nfl-draft-19702021

IMDb.com. (n.d.). *Draft day*. IMDb. Retrieved April 29, 2023, from https://www.imdb.com/title/tt2223990/plotsummary/?ref_=tt_ov_pl

Wikimedia Foundation. (2023, April 2). *Moneyball*. Wikipedia. Retrieved April 29, 2023, from https://en.wikipedia.org/wiki/Moneyball

Python Noteook:
https://colab.research.google.com/drive/1pkIHVnjq4bxL8QF05o0U5aWUb74K8QEl?usp=sharing