Investigating NBA Draftee Priorities Over Time

Team Name: Ashir Raza, Vincent Huang, Mina Kim, Elizabeth Berenguer

2020-11-11

Section 1: Introduction and Data

Our dataset consists of various statistics for every player who has played in the NBA within the 1996-2020 seasons, ranging from their home country and college of origin, draft year and round, and various stats such as average points, rebounds, and assists per game.

This dataset was found on Kaggle (https://www.kaggle.com/justinas/nba-players-data) and was originally collected using the NBA Stats API by Justinas Cirtautus, who created this dataset by filling in missing rows of data manually using data from the Basketball Reference website.

Each observation/case in the dataset represents a player and their corresponding qualities/ draft stats/game stats for a specific season. A player who played for 10 seasons within the analyzed timespan will have 10 observations. Some variables such as draft number remain constant, while others such as ppg change depending on the season. The variables in the dataset include different aspects pertaining to the player–including information about how/when they were drafted, what country/college they are from, physical characteristics (height, weight), and game stats (number of games played, rebounds).

Basketball is a constantly evolving game, and how NBA players played twenty years ago vs today is vastly different. From the rise of the three-point shooter to the fall of the big man, it is amazing how a game's rules can be so fluid over time. Our group sought to use this dataset to try and find more about how the highest flight of professional basketball has evolved over the years. Check the following article for more information: https://www.espn.com/nba/story/_/id/29113310/seven-ways-nba-changed-michael-jordan-bulls.

Research Question: How have the quality traits NBA teams desire in a draftee changed over time?

We will be looking at season to determine the time, the draft round and number to determine the relative importance of each player, and a way to combine a player's height, weight, and in-game statistical information in order to present an ideal position or skillset desired by teams for that draft year.

Our general hypothesis is that over the past twenty or so seasons, shorter and lighter players have been prioritized in drafts. Scoring, assisting, and true shooting averages will also likely have an upwards trend, while rebounding will likely trend downwards. Teams will have shifted from desiring a tall and heavy rebounder and shot-blocker towards a smaller and quicker shooter with more offensive capabilities.

Section 2: Methodology and Data

Taking a closer look at our data set by year, we can see that although there are some outliers, consistent data on players across the league began to be recorded around the year 1981. For this reason, we will pick 1996 as the starting year for our analysis and filter out any recorded data from before 1996. We also cap our analysis at before the 2019 season, as the 2019 season was ongoing at the time of upload, so metrics such as gp, pts, reb, etc. would not be complete.

Note that this data set is biased towards the offensive side of the ball, as it lacks defensive statistics such as blocks or steals. Therefore, the contributions of players focused on the defensive side may be neglected.

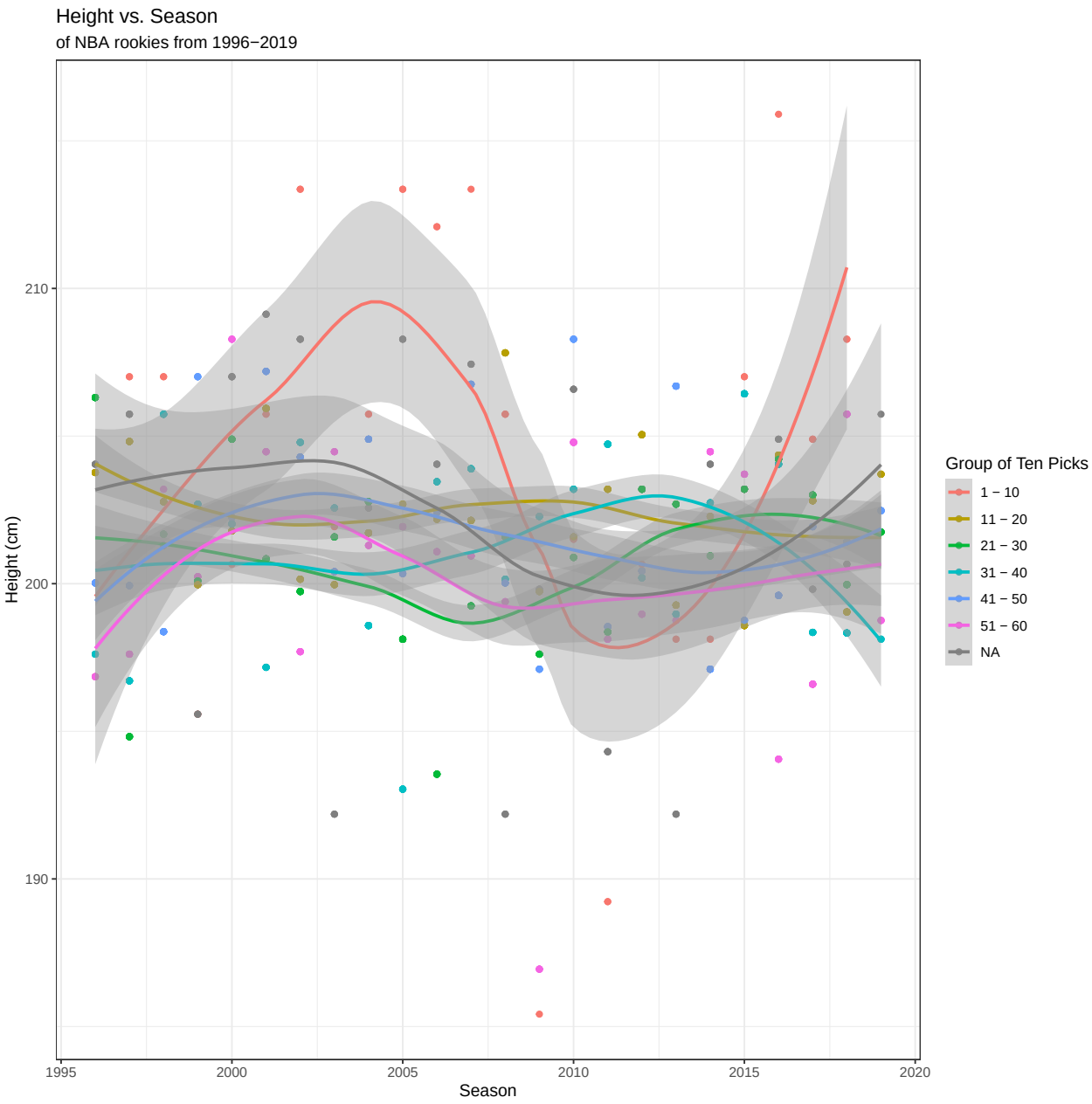
To begin, we create a new dataset called 'draft_data', which contains only newly drafted players. With the removal of all returning players, we are able to examine how drafting has changed from 1981 to 2018 and study what qualities NBA teams value over time. We essentially only took in the rookie seasons of players in

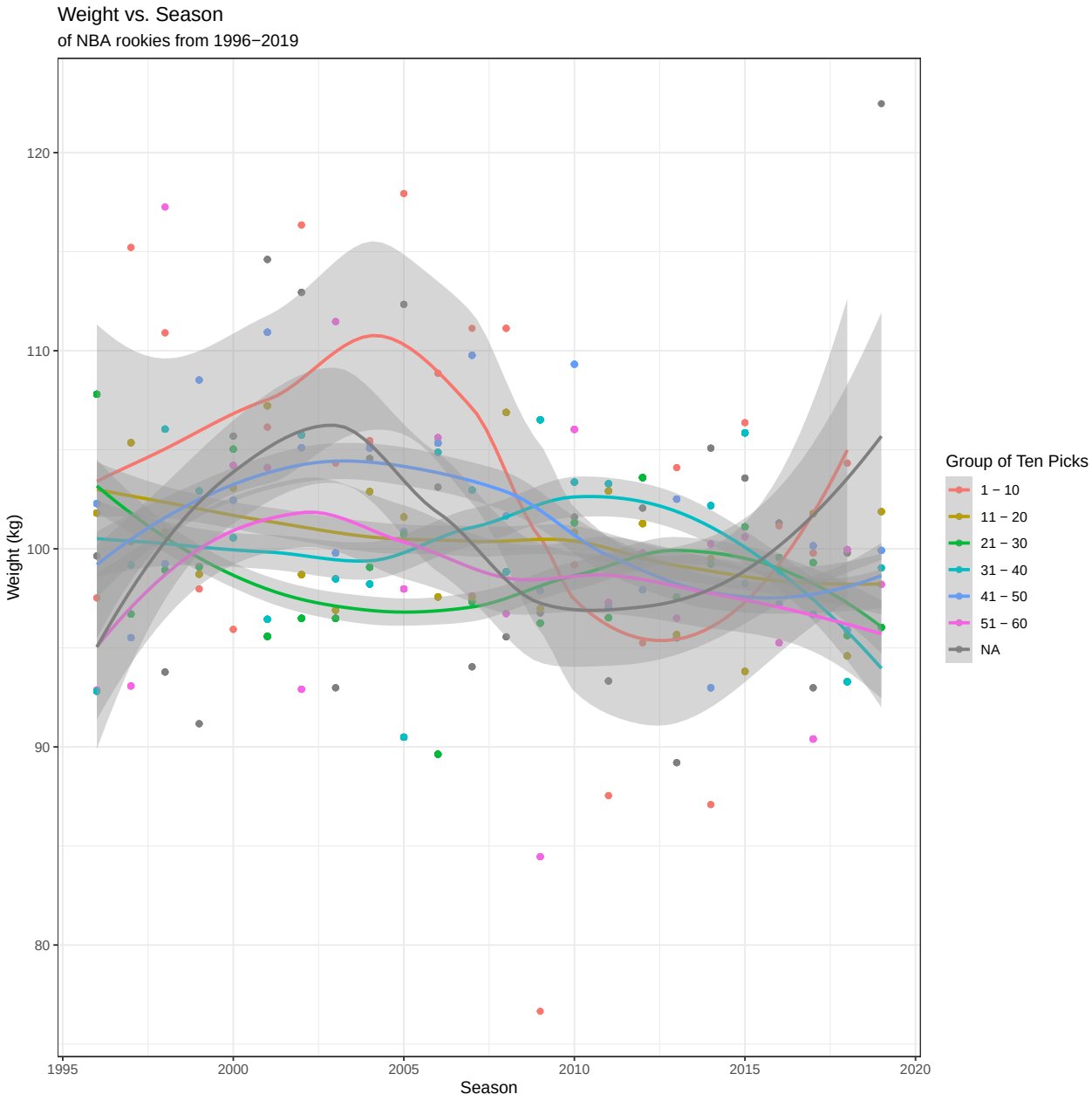
the dataset by filtering in only the observations that had a player's smallest age, which would correspond to their rookie season. We also remove data from the latest season since the information at the time of collection was still not complete due to the ongoing season.

Next, we group all of the draft picks by multiples of ten (ie: picks 1-10, 11-20, etc...) in order to organize the data more efficiently. These groups will be referred to as 'draft groups' for the remainder of this report.

We then group by both season and draft group, and find the mean of both height and weight–the two most obvious physical characteristics of an NBA player–for each season and draft group. We will graph these over time, with the draft groups differentiated by color, in order to determine if NBA teams' value of height and weight have changed at all over time.

We will be using scatterplots with smoothing lines to find how height and weight change over time, which we deemed to be the most appropriate graph because height and weight are both quantitative variables. It is important to note that although the season (our measurement of time) is technically stored as categorical information in the dataset, we can treat it as quantitative because R will order it numerically regardless. This is because we mutated the season variable to include just the first four characters, which is just the first year of the NBA season as NBA seasons run between two years.


Height vs. Season
of NBA rookies from 1996−2019
Height (cm)
210
200
190
1995
2000
2005
2010
2015
2020
Season
Group of Ten Picks
1 − 10
11 − 20
21 − 30
31 − 40
41 − 50
51 − 60
NA



Note that "NA" refers to players who were undrafted, but still picked up by an NBA team. Based on the two visualizations, it seems that both height and weight similarly changed over time. Both appeared to increase for the first few seasons, decrease for the next few seasons, and then increase once again for the latest few seasons from 1996 to 2019. This pattern was especially noticeable for picks 1-10, which are the most valuable picks. There appears to be no overall trend that occurs over the past twenty or so seasons for both weight and height.

Weight and height traditionally correlate with different positions in basketball, with the guards, or backcourt, being smaller and lighter and the forwards and centers, or frontcourt, being heavier and taller. To move forward, we will attempt to test how the value of both the frontcourt and the backcourt has changed over time, using statistical testing to determine how the selection and prioritization of guards has changed over time. We will be using both simulations of null distributions, and t-tests using the Central Limit Theorem, to find confidence intervals, more data visualizations, and most importantly, p-values.

Section 3: Results

Backcourt players or guards typically move the ball around frequently in order to deliver it to forwards or centers, who then score with the ball. They typically have many assists and fewer defensive rebounds, since they are usually located at the back of the court during an offensive play.

Thus, we have created a new variable called 'guards', which classifies if a rookie player is a guard or not. Our cutoff for this classification is 5 assists per game and a 20% defensive rebounding percentage. If a player has an average of 5 or more assists per game and a defensive rebounding percentage of 20% or less, then they are considered a guard. Otherwise, they are not a guard and are a forward or center.

We will test preferences for picking guards, and ask the following questions: Are more guards picked in the first draft round? Has the preference for guards shifted over time (is a greater proportion of guards picked in later seasons)?

Null Hypothesis: The proportion of guards picked in the first round is equal to the proportion of guards picked in the second round.

Alternate Hypothesis: The proportion of guards picked in the first round is greater than the proportion of guards picked in the second round.

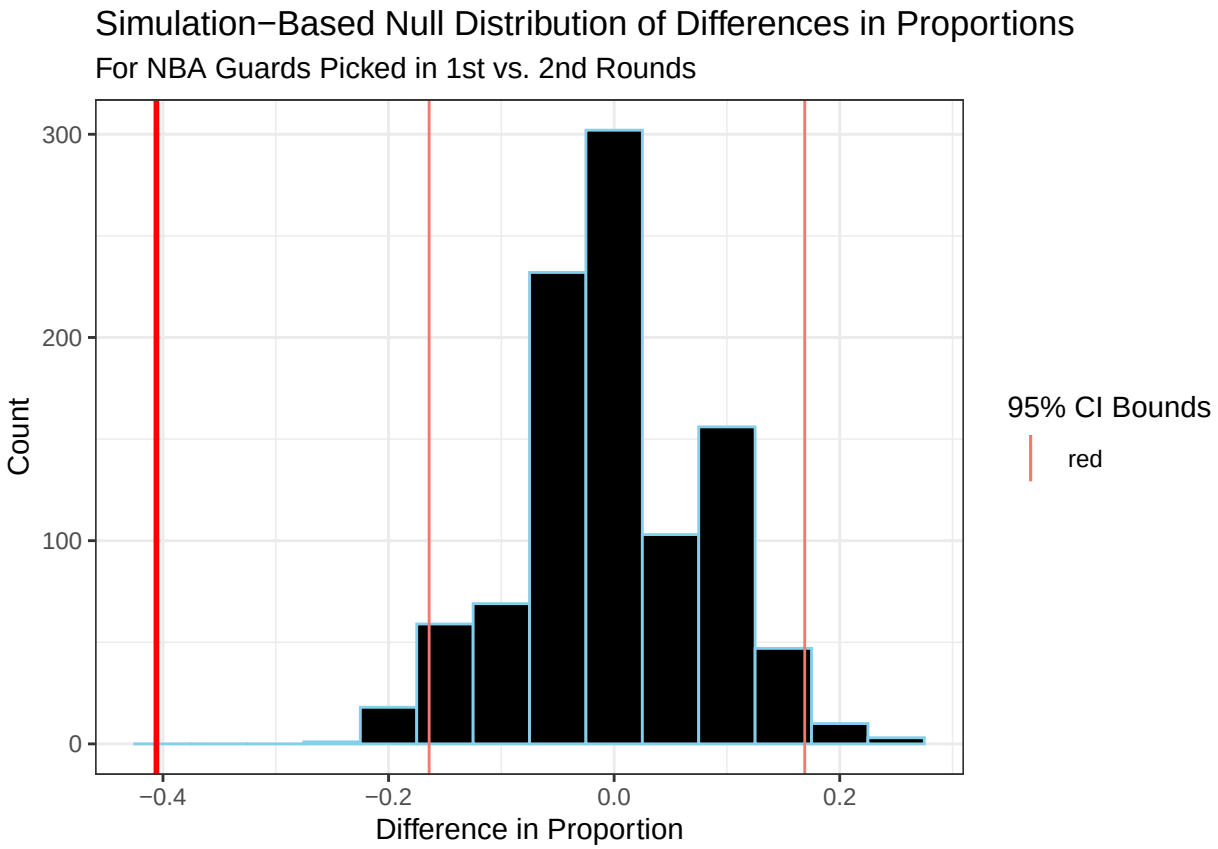
$H_0 : p_1 = p_2$

$H_a : p_1 > p_2$

To do this, we will use a simulation of a null distribution of difference in proportions. We will use 1000 samples and visualize them on a graph, along with a 95% confidence interval.

```
## # A tibble: 1 x 2
##    lower upper
##    <dbl> <dbl>
## 1 -0.164 0.169
```

```
## # A tibble: 1 x 1
##   p_val
##   <dbl>
## 1     0
```



Our p-value was about 0, and so we reject the null hypothesis at the a = 0.05 level. This decision is further supported by our 95% confidence interval, [-0.164, 0.169], and our null distribution graph, which shows that our observed difference is outside the confidence interval. There is sufficient evidence to suggest that a larger proportion of guards were picked in the first round compared to the second round, indicating that in general, skilled guards are likely more valued than forwards and centers.

We now test whether guards were consistently valued over the past 20 seasons, or if the players NBA teams valued were different in certain time periods. We can do this by comparing the mean of the season for both guards and frontcourt players to see which is higher or lower, if either. We will simulate a null distribution of differences in means, using 2500 samples.

Null Hypothesis: The mean year of guards who were drafted is not different from that of non-guards (frontcourt players) who were drafted.

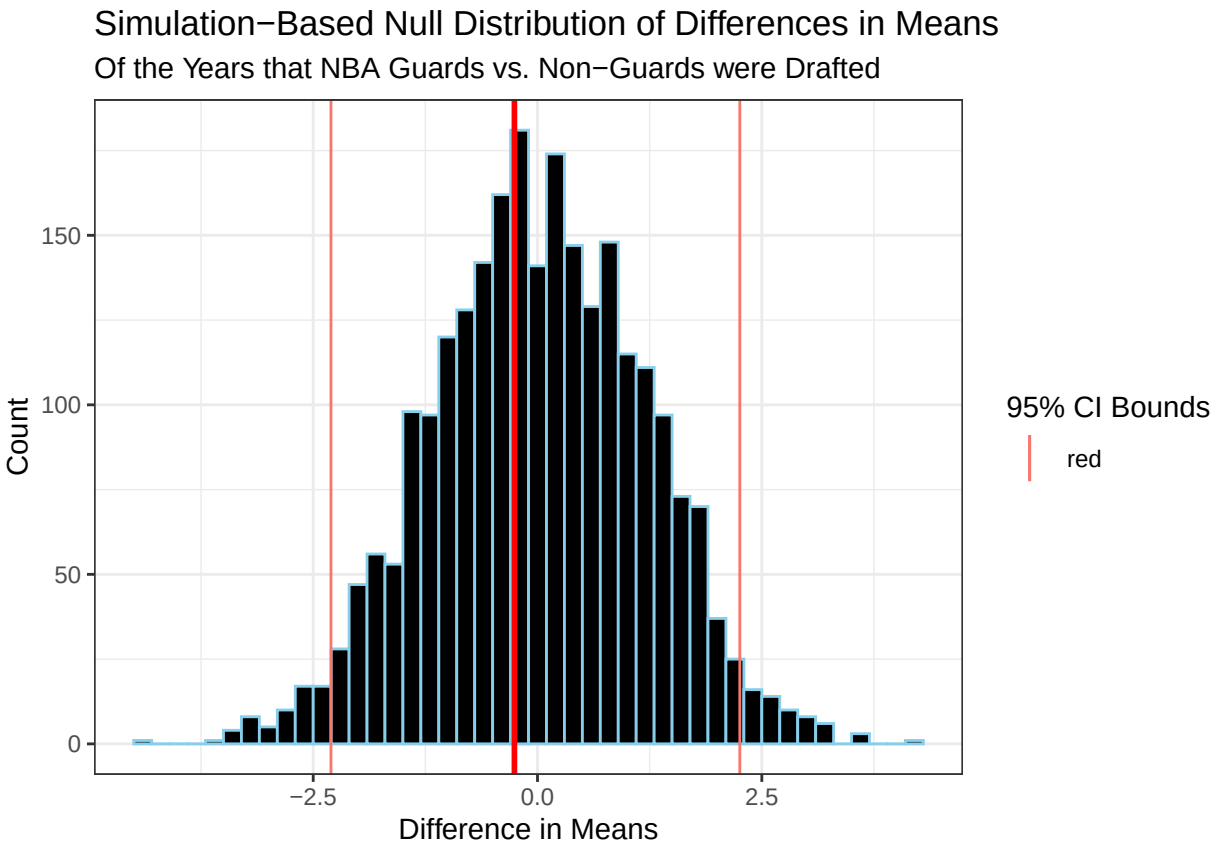
Alternative Hypothesis: The mean year of guards who were drafted is different from that of non-guards (frontcourt players) who were drafted.

$H_0 : \mu_1 = \mu_2$

$H_a : \mu_1 \neq \mu_2$

```
## # A tibble: 1 x 2
##   lower upper
##   <dbl> <dbl>
## 1 -2.30  2.26
```

```
## # A tibble: 1 x 1
##   p_val
##   <dbl>
## 1 0.585
```



Our p-value was 0.5848, which is much greater than a = 0.05, and so we fail to reject our null hypothesis. There is not sufficient evidence to suggest that the mean year of guards who were drafted is different from that of non-guards (frontcourt players) who were drafted. In the context of our research question, there is not sufficient evidence to suggest that NBA teams' preferences have shifted towards picking guards during the first draft round.

Because our findings have so far not been conclusive, we will now attempt to analyze whether frontcourt players were more valued in certain seasons based on our visualizations in Section 1.

From our height and weight graphs, we observed that there was a stronger preference for players who were taller and weighed more in the early 2000s (from about 2000 to 2006). We will test whether this preference for taller and heavier players influences the number of rebounds, using the Central Limit Theorem.

We will use a t-test on data that is both unaltered and data that includes only the 2000-2006 seasons. If the subset has higher interval ranges, then that would mean the rebound mean would be higher in those 6 seasons, suggesting that taller and heavier players were likely more valued during that time period.

```
## # A tibble: 1 x 6
##   statistic  t_df   p_value alternative lower_ci upper_ci
##       <dbl> <dbl>     <dbl> <chr>          <dbl>    <dbl>
## 1      47.3  1197 1.96e-276 two.sided       2.48     2.69
```

Using a two-sided t-test with a test statistic of 47.329 and 1197 degrees of freedom, we found a p-value of 0 and a confidence interval of [2.478, 2.692]. We are 95% confident that the mean number of rebounds for players falls between 2.478 and 2.692. If we were to repeatedly and independently sample from the original population and construct 95% CIs in the same way, we would expect 95% of such intervals to truly contain the population mean.

```
## # A tibble: 1 x 6
##   statistic  t_df  p_value alternative lower_ci upper_ci
```

```
##       <dbl> <dbl>    <dbl> <chr>          <dbl>    <dbl>
## 1      26.3   355 2.39e-85 two.sided       2.28     2.65
```

Using a two-sided t-test with a test statistic of 26.296 and 355 degrees of freedom, we found a p-value of 0 and a confidence interval of [2.281, 2.65]. We are 95% confident that the mean number of rebounds for players from the early 2000s (years 2000-2006) falls between 2.281 and 2.65. If we were to repeatedly and independently sample from the original population and construct 95% CIs in the same way, we would expect 95% of such intervals to truly contain the population mean.

However, our confidence interval ranges are lower for the dataset with only the 2000-2006 seasons, implying that taller and heavier players are not as valued since confidence intervals and p-values have a 1 to 1 correspondence.

Section 4: Discussion and Conclusion

We learned that taller and heavier NBA players were not more preferred in the 2000-2006 seasons, despite the implications of our height and weight graphs over seasons, categorized by draft group. We reached this conclusion through our CLT test results, which showed that the 95% confidence interval for the early 2000s, [2.28, 2.65], had lower values than that of the overall data set, which was [2.48, 2.69].

We also learned that the proportion of guards picked in the first round was greater than that of guards picked in the second round. Because our p-value for the test we used was 0, which was lower than our significance level of 0.05, we concluded that guards are generally prioritized by NBA teams during the draft process.

However, we also learned that there is no relation between prioritizing guards and frontcourt players over the years. Because our p-value for the test we used was 0.634, which was much larger than our significance level of 0.05, we reached the conclusion that NBA teams tend to prefer guards over frontcourt players, but also that drafting patterns have generally been similar over the past 20 seasons.

Our data visualization methods in Section 1 were not as reliable or precise as we would have hoped. Although we were able to map out draft picks over twenty years groups by draft groups, the margins of error were fairly large as seen by the shaded regions around our smoothing lines. We might address this issue by creating visualizations for several seasons at a time, rather than trying to represent the entire dataset with one graph.

Our CLT t-tests were also lacking in validity and reliability. We did not compare the rebounding averages to another statistic, and only compared the two confidence intervals in order to make our conclusion. This is likely not very accurate since we did not use p-values. However, since confidence intervals and p-values have a 1 to 1 correspondence, our analysis should still be somewhat correct, although a proper simulation or a more detailed CLT t-test would likely be more accurate.

Our null distribution simulations are probably the most accurate and reliable part of our statistical analysis. To improve in this aspect, we might increase the number of samples per simulation.

If we were to redo this analysis, we would try to filter our data more. Even after we filtered out every non-rookie player in our large dataset, we still had about 1200 observations, which is large enough to increase the probability of errors and outliers. We would also try to facet our data and then graph it to facilitate ease of analysis and interpretation.

In the future, we want to analyze age and how the age of NBA draftees has–or has not–changed over time. The debate between going pro as soon as possible or staying four years at a Division 1 college is a major point of contention, and it would be an interesting question to explore.

We could also have more inference testing related to frontcourt positions, since we focused on a lot more than intended on the backcourt for this lab report. Many famous players right now such Lebron James or Kevin Durant are all frontcourt players, so it would be interesting to explore that particular aspect of the NBA.