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 ingame 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 probably have a downwards trend. 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

If we take 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 started being kept around the year 1981. For this reason, we will pick 1996 as our start year for our analysis and filter out any recorded data from before 1996. We will also cap off our analysis before the 2019 season, as that season was ongoing at the time of upload, so metrics like gp, pts, reb, etc. would not be complete.

This data set is biased towards the offensive side of the ball, with there being no defensive statistics such as blocks or steals in this data set, so contributions given by players who devoted much of their time to the defensive side of the ball might be neglected.

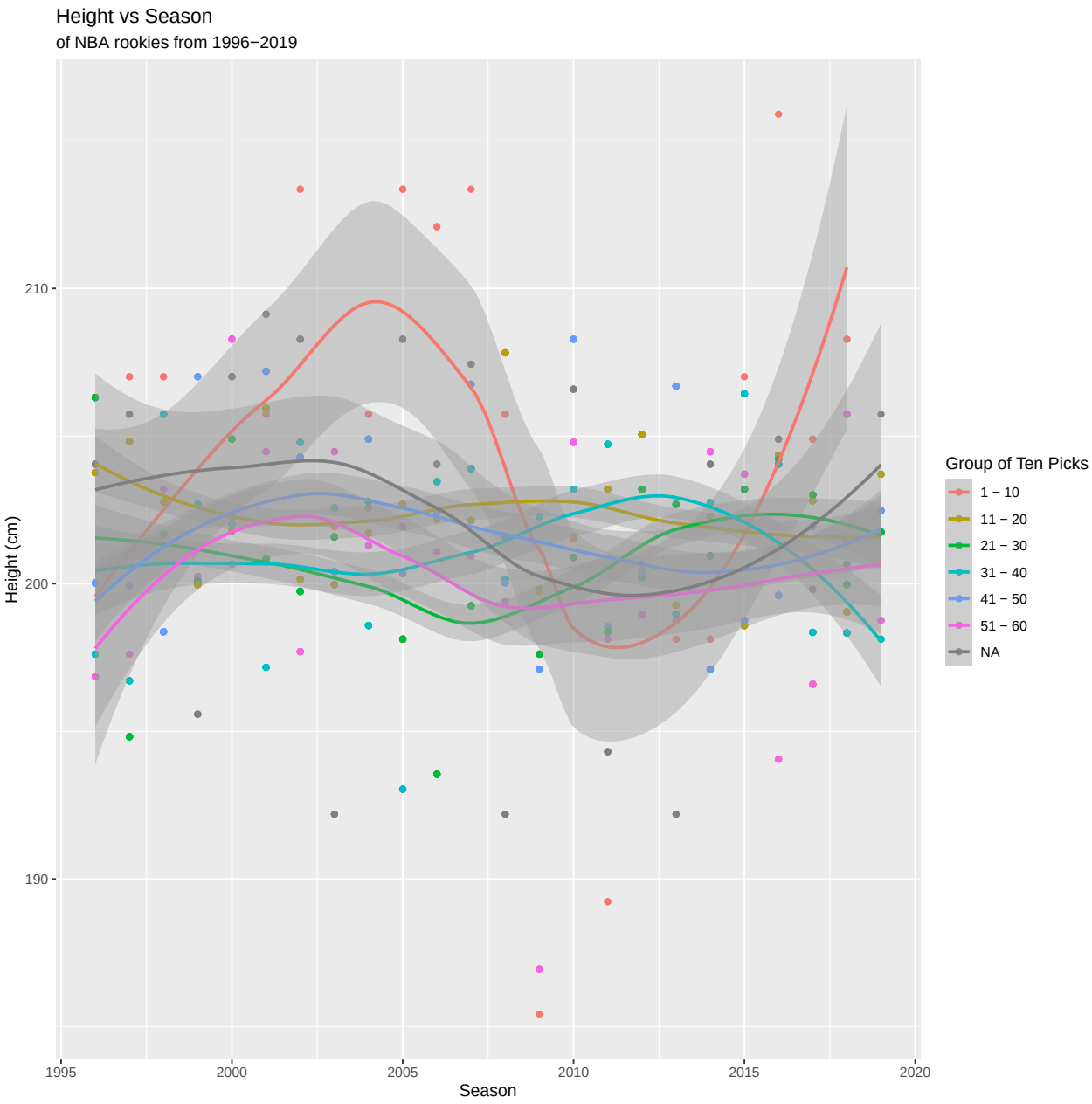
To start, we will create a new dataset with just newly drafted players called draft_data. This will remove all returning players to allow us to simply look at how drafting has changed from 1981 to 2018 and to try to

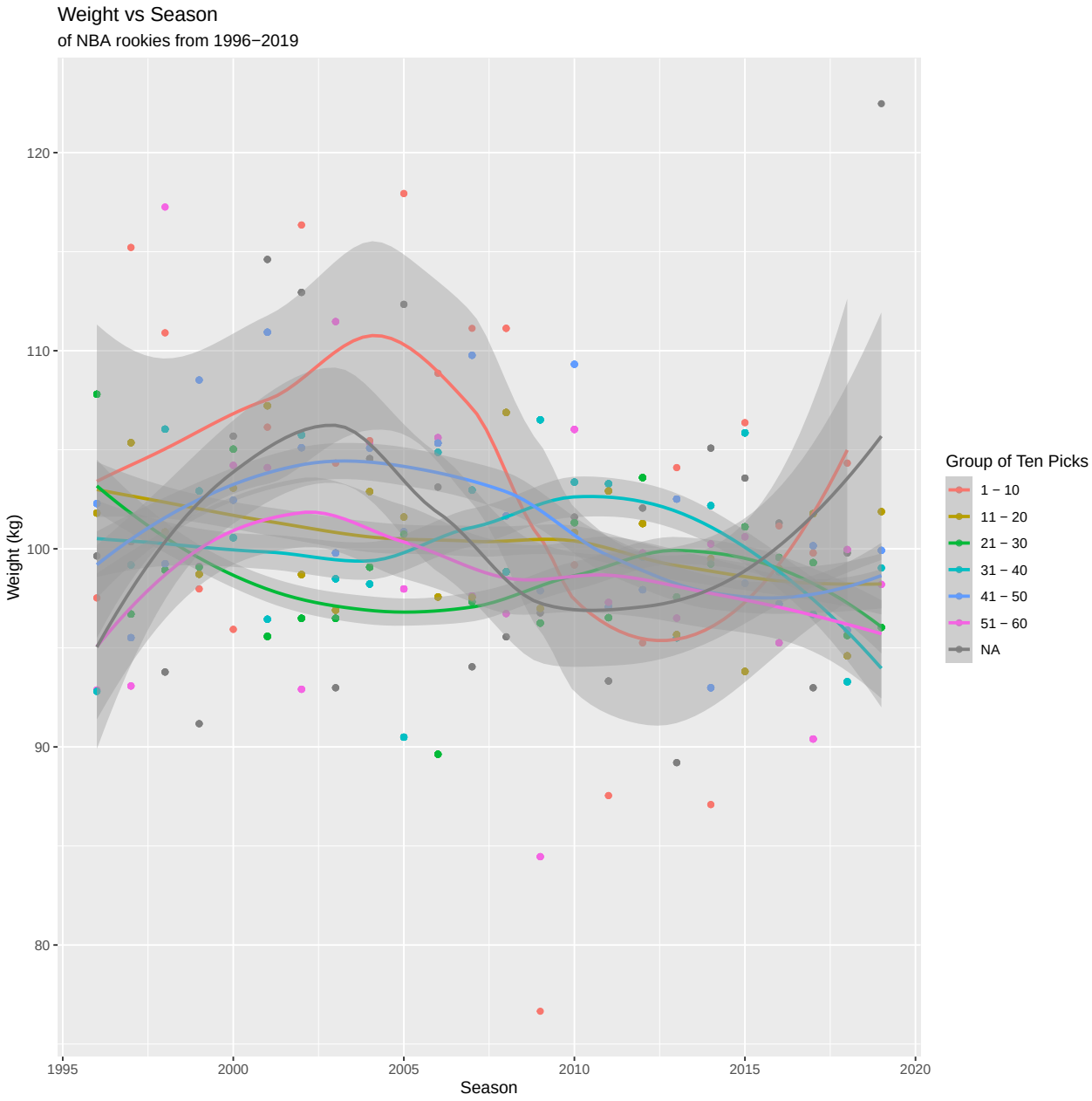
view what qualities NBA teams value over time. We essentially only took in the rookie seasons of players in the dataset by filtering in only the observations that had a player's smallest age, which would correspond to their rookie season. We also removed data from the latest season since the info at the time of collection was still not complete as the season was still ongoing.

We shall also group all 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 shall now 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, which we deemed to be the most appropriate graph due to the fact that we are trying to find how both height and weight change over time, which is comparing two 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 since R will order it numerically anyway. 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



For further information, "NA" refers to players who were undrafted but still picked up by an NBA team. Based off of the two visualizations, it seems that both height and weight similarly changed over time. Both appeared to increase for the first few seasons and then decrease for the next few seasons, and then increase once again for 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 shall attempt to test how the value of both the frontcourt and the backcourt has changed over time.

To move on in this investigation, it would be a wise idea to use 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-testing using 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 a lot in order to deliver it to forwards or centers who then score the ball. They typically have lot of assists and fewer defensive rebounds since they are typically located at the back of the court during an offensive scheme.

Thus, we have created a new variable called guards that classifies if a rookie player is a guard or not. Our parameters for this are 5 assists per game and 20% defensive rebounding percentage. If a player has an assists per game average of 5 or greater 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: 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 to the proportion of guards picked in the second round.

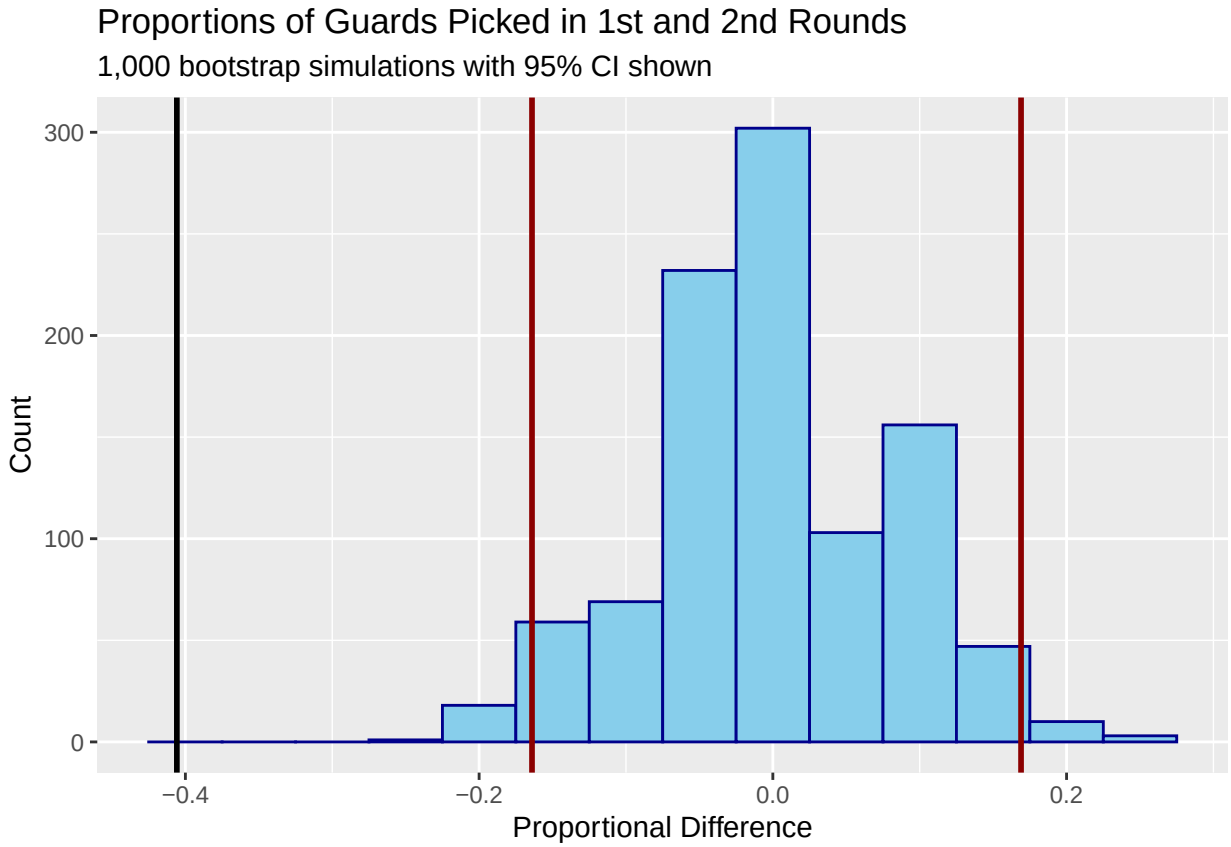
$H_0 : p_1 = p_2$

$H_a : p_1 > p_2$

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 is further backed by our 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 guards of guards were picked in the first round compared to the second round, meaning that skilled guards in general are likely more valued than forwards and centers.

We shall now test whether guards were consistently valued over the past twenty seasons, or if certain time periods had different values in what NBA teams wanted in a player. 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.

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

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

$H_0 : \mu_1 = \mu_2$

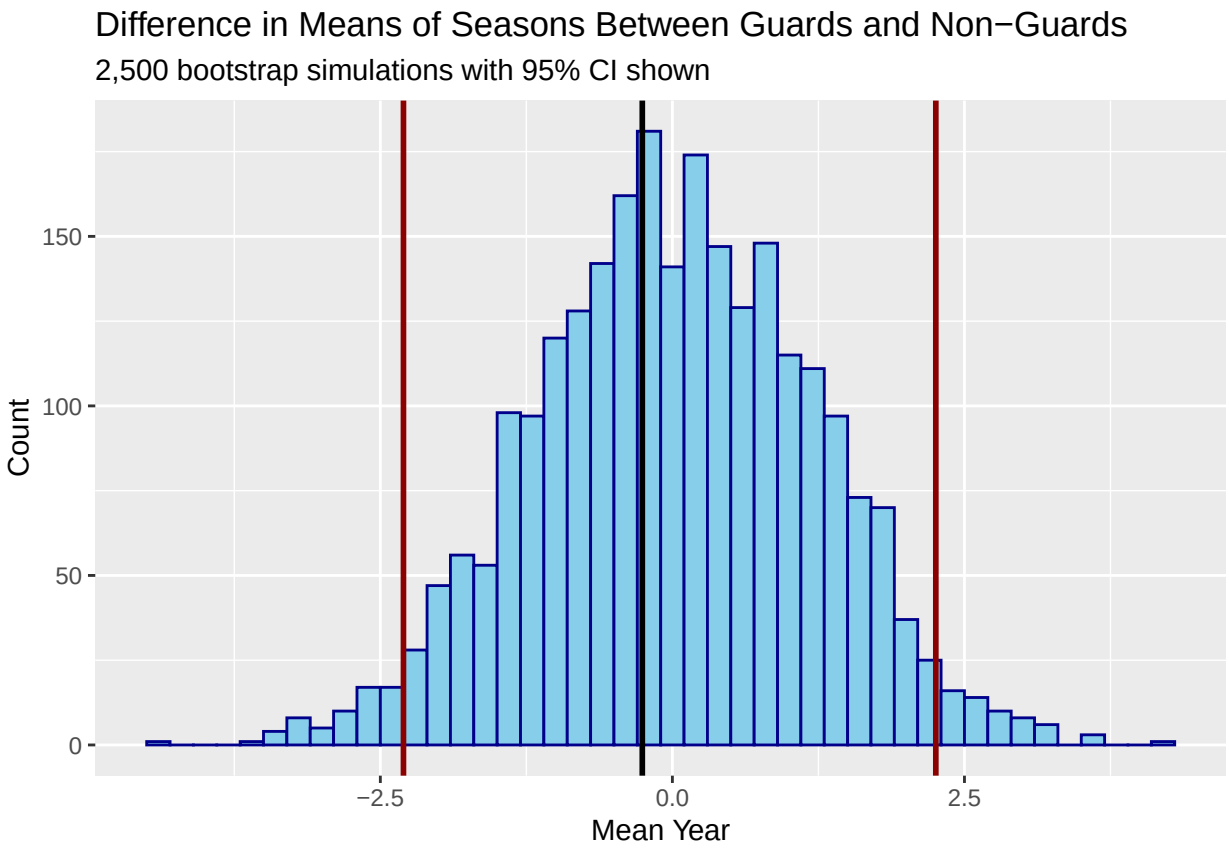
$H_a : \mu_1 \neq \mu_2$

Similar to the previous test, we will use a simulation of a null distribution. However, this time, it will be of a difference in means. We will use 2500 samples and visualize them on a graph, along with a 95% confidence interval.

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

```
## # A tibble: 1 x 1
##   p_val
```

```
##   <dbl>
## 1 0.634
```



Our confidence interval was [-2.302, 2.255]. Our null distribution graph shows that our observed difference in means is approximately in the middle of the confidence interval, which shows that we will likely not be able to reject our null hypothesis.

This is proven by our p-value, 0.634, which is much greater than 0.05, so we fail to reject our null hypothesis. Therefore, there is no sufficient evidence that the mean year of guards who were drafted is either different or not different than non guards who were drafted. In the context of our research question, there is not sufficient evidence to prove that there has been a shift in preferences to prefer picking guards during the first draft round.

In order to move on from our disheartening lack of results, we will attempt to analyze whether frontcourt players were more valued in certain seasons based off our visualizations in Section 1.

From our height and weight graphs, we saw that there was a larger 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 is filtered for just 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, meaning that taller and heavier players are 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 is between 2.478 to 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) is 2.281 to 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 did not have a higher preference from the 2000-2006 seasons, even though it was implied by our visualization graphs of height and weight over the seasons and categorized by draft group. This was shown by 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. We know this from our simulation null distribution, which had a p-value of about 0, which is lower than the 0.05 alpha level we tested it on. Thus we know 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. We know this from our other simulation null distribution, which had a p-value of 0.634, which is much bigger than the 0.05 alpha level we tested our statistics on. Overall, we learned about our research question that NBA teams generally prefer guards over frontcourt players, but that drafting patterns have generally similar over the past 20 seasons.

Our data visualization methods in Section 1 could use some improvement. 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. This proves that our visualization was not completely reliable or valid, which is definitely a problem. This could probably be fixed through having several visualizations over a few seasons at a time rather than trying to represent the entire dataset within one graph.

Our CLT t-tests are also probably lacking in validity and reliability. We did not use proper inference procedures and compare the rebounding averages to another statistic, and only compared the two confidence intervals in order to make our conclusion. This is probably not very accurate since we did not use any p-value, even though confidence intervals and p-values have a 1 to 1 correspondence. Due to that, our analysis should still be somewhat correct, but a proper simulation or a more detailed CLT t-test would probably be more accurate.

Our null distribution simulations are probably the most accurate and reliable part of our statistical analysis. The only thing we could improve at in this aspect would be to increase the number of samples per simulation.

Something we would do differently if we were to redo this experiment would be to try to filter out our data more, since even after we filtered out every non rookie player in the massive dataset, we still had about 1200 observations, which is a massive number that would be prone to errors and outliers.

We could also try to facet our data and then graph it instead of having the congested visualization graphs we currently have in Section 1. This would make our data easier to read and analyze.

Something we would do in the future for this experiment would be to analyze age and how the age of draftees has or has not changed over time. The debate between going pro as soon as possible or staying four years at a D1 college is a major sports debate right now, 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 another interesting aspect of this topic to explore.