Investigating NBA Draftee Priorities Over Time

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

2020-10-29

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. Some variables such as draft number remain constant while other such as ppg change depending on the season. A player who played for ten seasons within the timespan analyzed will have ten observations. The variables in the dataset include different aspects pertaining to the player- whether it be information about how/ when they were drafted, what country and 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 vary tremendously. From the rise of the three point shooter to the fall of the big man, it's amazing how a game's rules can be so fluid over time. Our group wishes 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.

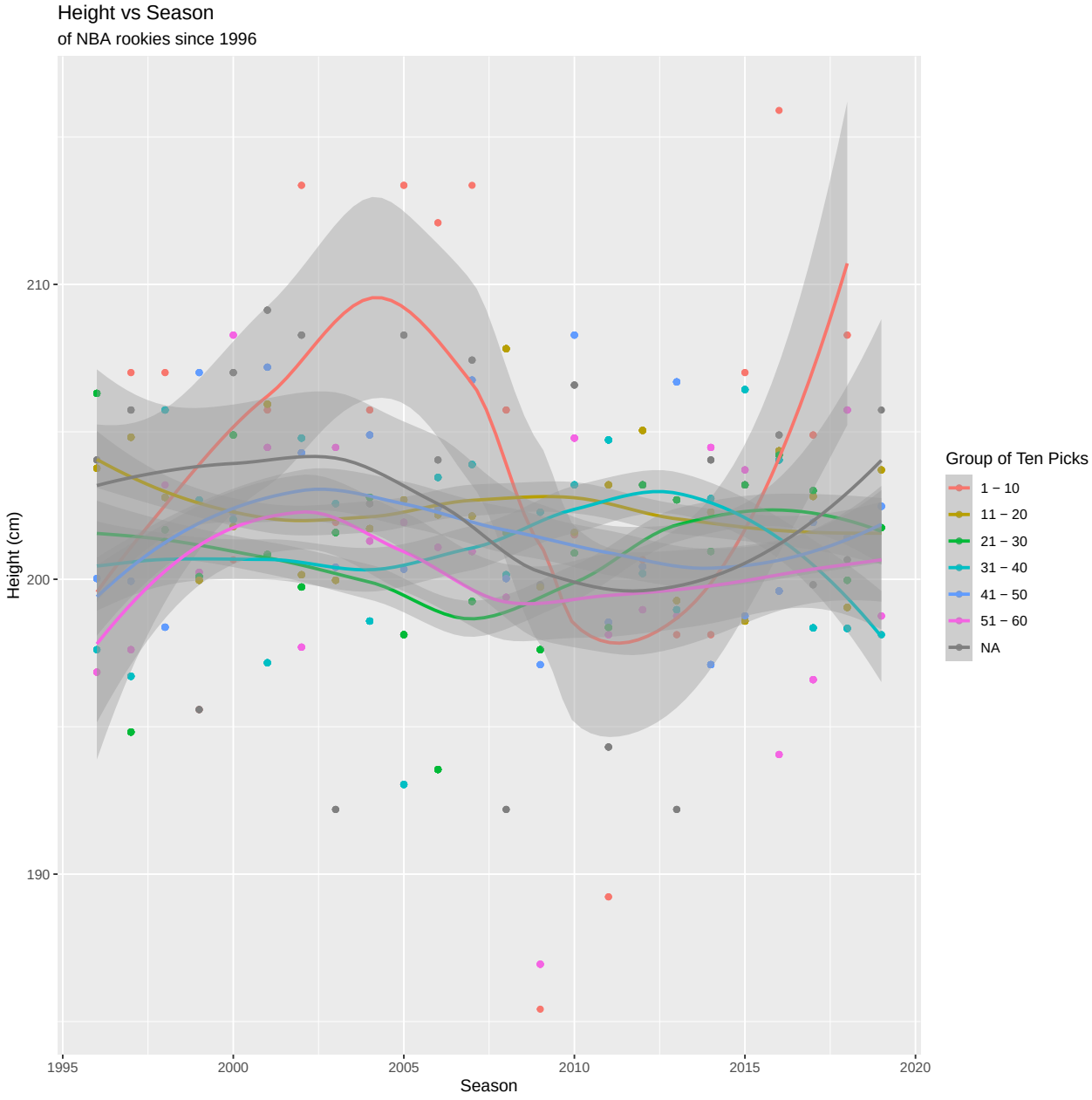
This data set is biased towards the offensive side of the ball, with there being no defensive statistics such as blocks or steals in this data set, so contributions given by players who devoted much of their time to the defensive side of the ball might be neglected.

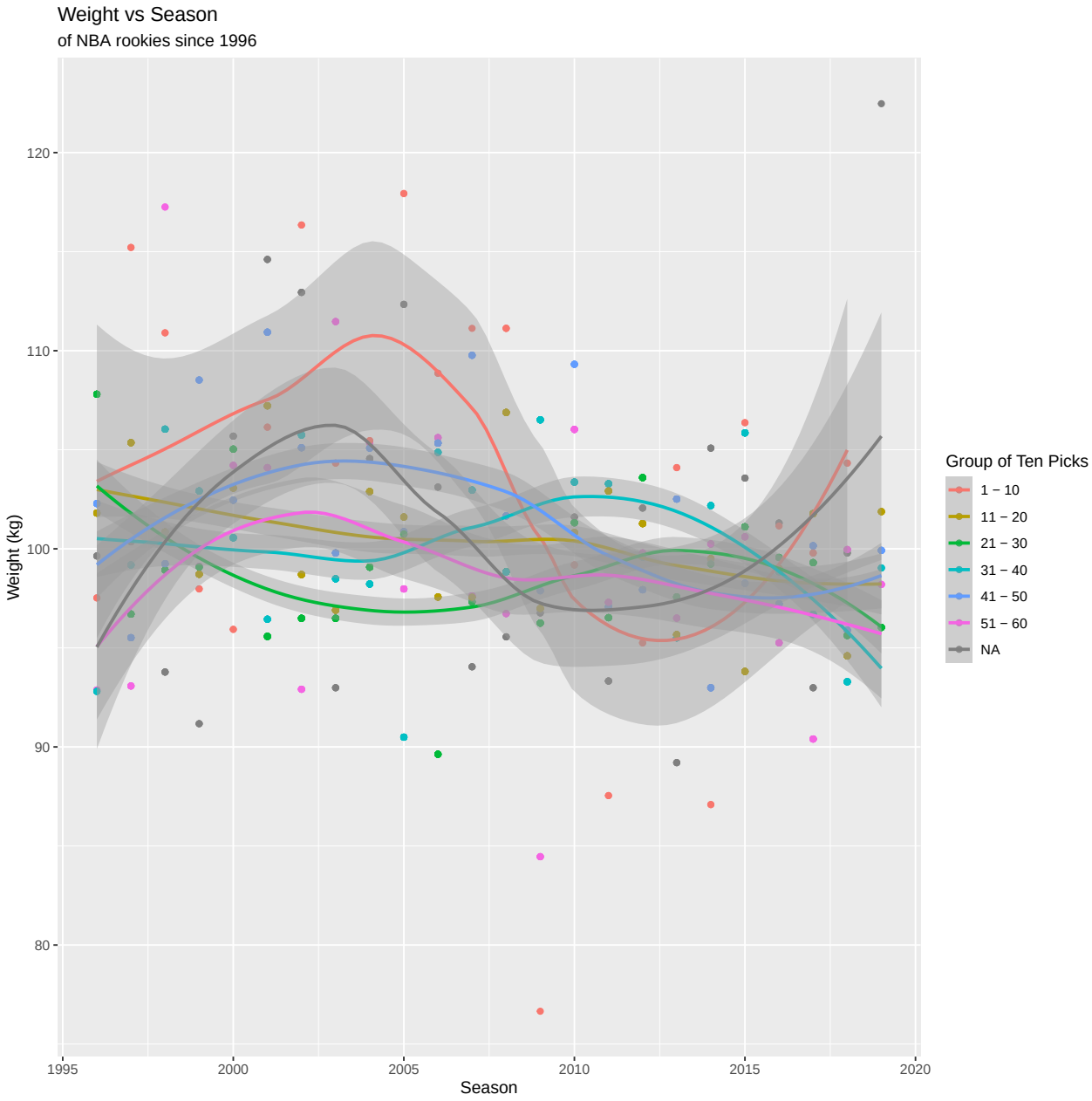
To start, we will create a new dataset with just newly drafted players. This will remove all returning players to allow us to simply look at how drafting has changed from 1981 to 2018 and to try to view what qualities

NBA teams value over time. We 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.





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

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

H0: The proportion of guards picked in the first round is equal to the proportion of guards picked in the second round.

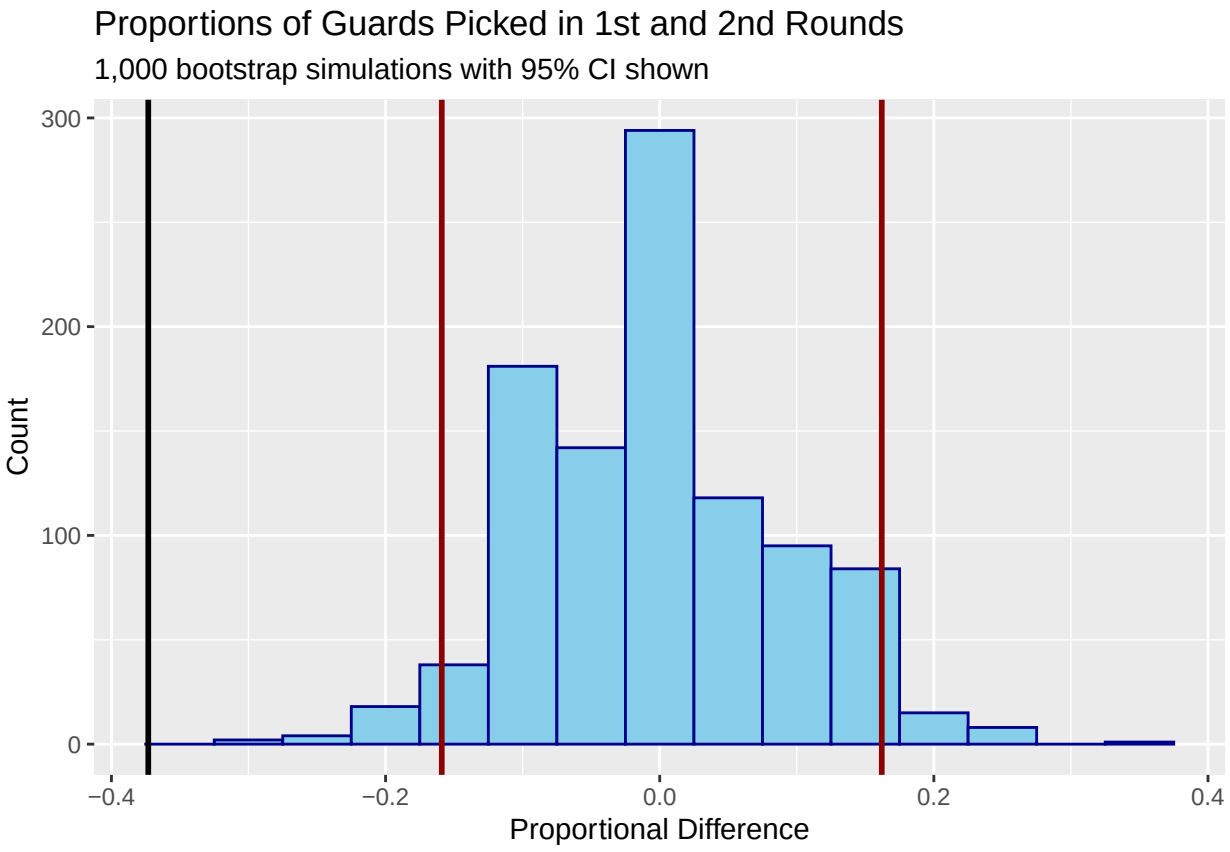
Ha: The proportion of guards picked in the first round is greater to the proportion of guards picked in the second round.

$H_0 : \chi_1^2 = \chi_2^2$

$H_a : \chi_1^2 > \chi_2^2$

```
## # A tibble: 1 x 2
##    lower upper
##    <dbl> <dbl>
## 1 -0.159 0.162
```

```
## [1] 0
```



Our p-value was about 0, and so we reject the null hypothesis at the a = 0.05 level. There is sufficient evidence to suggest that a larger proportion guards were picked in the first round compared to the second round.

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

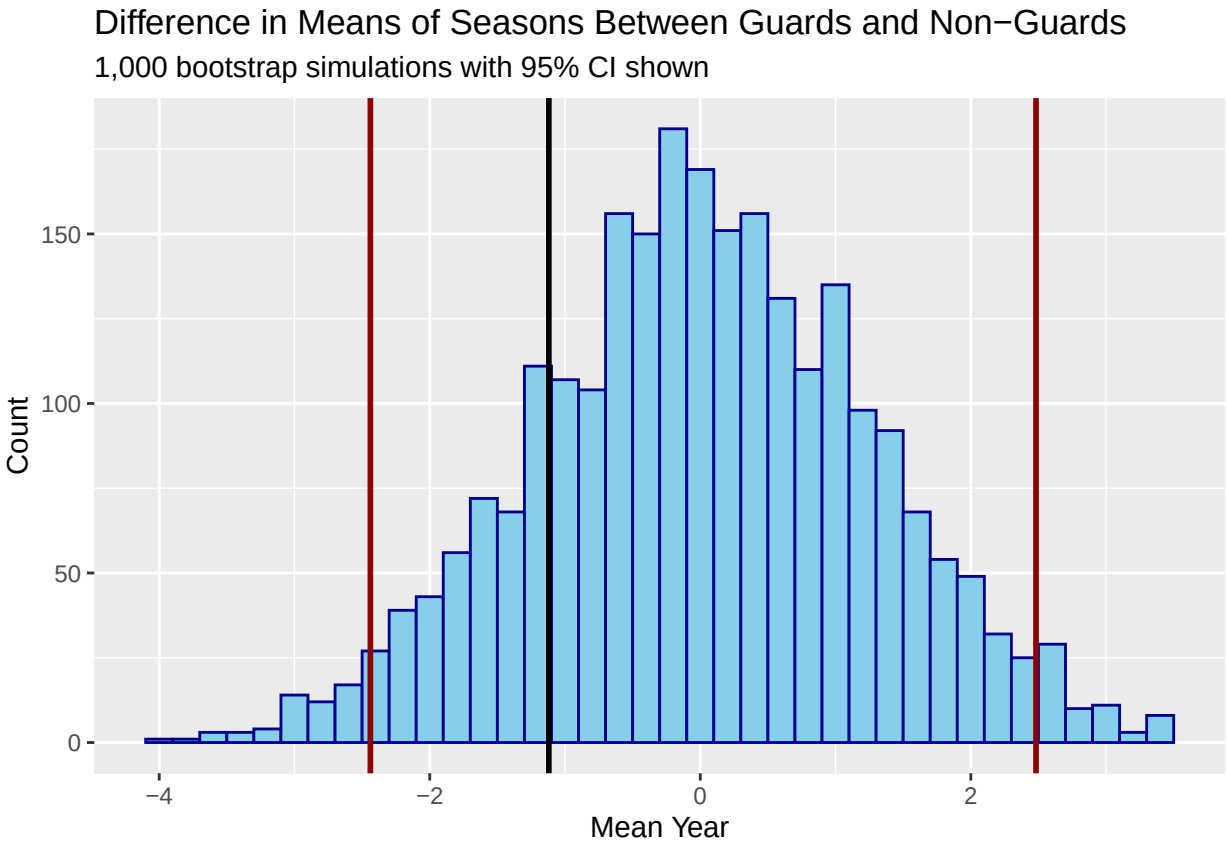
Alternative: The mean year of guards who were drafted is different than that of non-guards (frontcourt players) who were drafted.

$H_0 : \mu_1 = \mu_2$

$H_a : \mu_1 > \mu_2$

```
## # A tibble: 1 x 2
##   lower upper
##   <dbl> <dbl>
## 1 -2.44  2.48
```

```
## # A tibble: 1 x 1
##   p_val
##   <dbl>
## 1 0.626
```



Since our p-value, 0.8116, is much greater than 0.05, we fail to reject our null hypothesis. Therefore, there is no sufficient evidence that the mean year of guards who were drafted is 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.

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 player influences the number of rebounds.

```
## # 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
```

We are 95% confident that the mean number of rebounds for players is 2.477 to 2.691 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
```

We are 95% confident that the mean number of rebounds for players from the early 2000s (years 2000-2006) is 2.281062 to 2.649837 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.

Section 4) Discussion and Conclusion