Data Exploration Project Report

Tung X. Nguyen

June 3, 2020

Contents

List of Figures

1 Introduction

In most types of sports, physicality is an important indicator of players for their potential to success. This is especially true for basketball, which is a very physical game. Before 1990, basketball players were mostly tall, sometimes skinny men. As time goes by, we see more players with more body mass. This undeniable trend is a proof showing that physique in NBA is gaining more importance.

Inspired by this observation and by my interest in the game, I decided to pick ”The importance of physique in NBA” as the topic for this report. In this report, I attemp to answer several key questions:

- What is the most common physique type in the NBA (from 2000 to 2020) by height and weight?
- How the distribution of height and weight changes?
- How physique can affect the draft number of a player?
- Is there any differences in the physique of non-US players to US players?
- How physique and Player Efficiency Rating correlated?

For the second question, I split the data into 2 groups for comparison: drafted before 2002 and drafted after 2002. For the third question, I use all draft data from 2000 to 2019 to answer. For the final question related to PER, I only use the PER data from two season 2005-2006 and 2018-2019 for comparison.

2 Data Wrangling

The subsection **Data sources** belows shows the description of my sources. All of them are tabular, html data that can be scraped with the help of libraries including Selenium, BeautifulSoup, pandas and csv.

2.1 Data sources

2.1.1 NBA players regular season stats from Official NBA Statistics and Advanced Analytics

Link: *stats.nba.com/players/bio/*

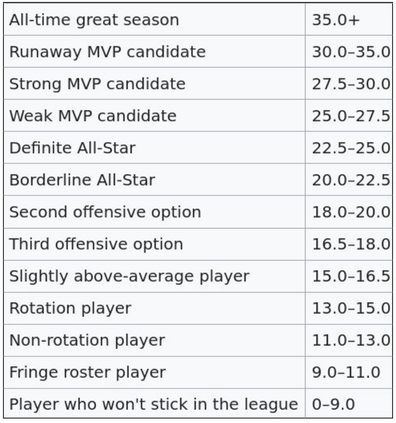
This website contain the information about players' full name, their age, their draft, height and weight. There is also the information about their draft (including draft year, draft round, draft number) and key perfomance measurement and the team they are playing for, and these information usually changes season after season. The number of players varies each season, but the maximum is 450 (15 players a team multiplied by 30 teams).

2.1.2 Player Efficiency Rating (PER)

Link *insider.espn.com/nba/hollinger/statistics.*

This is a scale developed by John Hollinger, former Vice President of Basketball Operations for the Memphis Grizzlies (an NBA team). This all-in-one formula attempts to calculate a player's contribution per playing minute, taking in consideration key performance items such as field goals, assists, steals, blocks, rebounds, free-throw, three-pointers (Wikipedia contributors, 2020). PER is not the perfect scale to measure a player defensively (as good defensive players are not necessarily excellent blockers or stealers), but it is still one of the most popular tools available to do evaluate players. Figure 1 is a reference table for PER, provided by Hollinger himself.

Figure 1: PER reference guide



All-time great season	35.0+
Runaway MVP candidate	30.0–35.0
Strong MVP candidate	27.5–30.0
Weak MVP candidate	25.0–27.5
Definite All-Star	22.5–25.0
Borderline All-Star	20.0–22.5
Second offensive option	18.0–20.0
Third offensive option	16.5–18.0
Slightly above-average player	15.0–16.5
Rotation player	13.0–15.0
Non-rotation player	11.0–13.0
Fringe roster player	9.0–11.0
Player who won't stick in the league	0–9.0

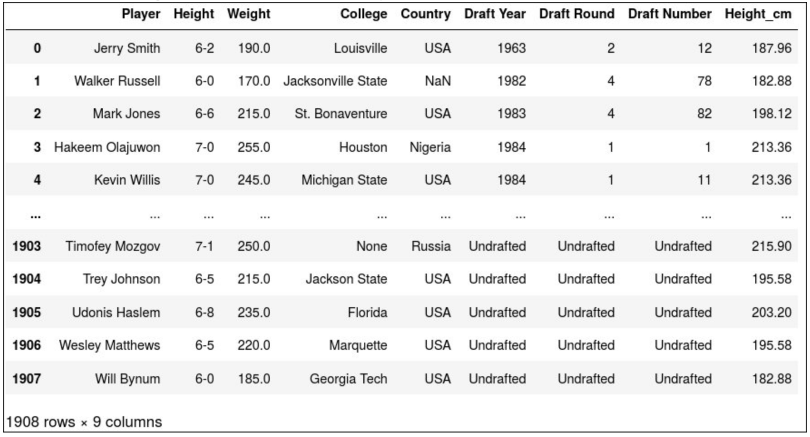
The PER dataset I used for this assignment is for regular seasons only. Available data includes: the number of games played, minutes per game, true shooting percentage, assist ratio, turnover ratio, usage rate, offensive/defensive rebound rate, PER, value aded, and estimated wins added. Explanation and formula for each of these columns can be found in the end of the data table on the website.

2.2 Data Cleaning and Transformation

After the data is saved to csv files, I proceed to transform them into the format I wanted.

For the NBA players regular season data, I retained only the columns that contain data about physique and draft. Next, I perform union on datasets across seasons from 2000-2001 to 2019-2020 to finally get the data of total 1965 players playing in the NBA from 2000 to 2020. I then remove 57 entries that contains null in physique information. The **Height** column is originally in feet-inch format, so I used regex to extract tokens and transform this column in to a new column named **Height_cm**. **Weight** is in pound unit. The result is a dataframe like in Figure 2. The dataframe is sorted according to the **Draft Year** and **Draft Round** columns. It can be seen that there are many players who were undrafted, but I decided to keep all these entries (there are 562 undrafted players).

Figure 2: NBA players from 2000 to 2020



	Player	Height	Weight	College	Country	Draft Year	Draft Round	Draft Number	Height_cm
0	Jerry Smith	6-2	190.0	Louisville	USA	1963	2	12	187.96
1	Walker Russell	6-0	170.0	Jacksonville State	NaN	1982	4	78	182.88
2	Mark Jones	6-6	215.0	St. Bonaventure	USA	1983	4	82	198.12
3	Hakeem Olajuwon	7-0	255.0	Houston	Nigeria	1984	1	1	213.36
4	Kevin Willis	7-0	245.0	Michigan State	USA	1984	1	11	213.36
...	...	...	...	...	...	...	...	...	...
1903	Timofey Mozgov	7-1	250.0	None	Russia	Undrafted	Undrafted	Undrafted	215.90
1904	Trey Johnson	6-5	215.0	Jackson State	USA	Undrafted	Undrafted	Undrafted	195.58
1905	Udonis Haslem	6-8	235.0	Florida	USA	Undrafted	Undrafted	Undrafted	203.20
1906	Wesley Matthews	6-5	220.0	Marquette	USA	Undrafted	Undrafted	Undrafted	195.58
1907	Will Bynum	6-0	185.0	Georgia Tech	USA	Undrafted	Undrafted	Undrafted	182.88

1908 rows × 9 columns

In the similar manner, I retrieve the PER data for each regular season. I then join (using Python) these data with the physique data above.

3 Data Checking

There are some missing in my dataset. Firstly, there are Guy Rucker and Xavier Silas, whose draft number is leave empty in the original dataset. After looking for their information online, I change their draft number to Undrafted.

Next, after merging the PER data with the physique data, I found out that there are some players who were not recorded in the physique data but actually played at the momment. So, I looked up online for their height and weight information to manually fill in the data, using Python.

4 Data Exploration

In this section, I use Rstudio to draw plots and explore data.

4.1 The distribution of Height and Weight

The distribution of **Height** for all players from 2000 to 2020 is shown in Figure 3. This distribution is skewed to the right, and the most popular height range is from about 195 to 210 centimeters.

Figure 3: Density plot of Height (in cm)

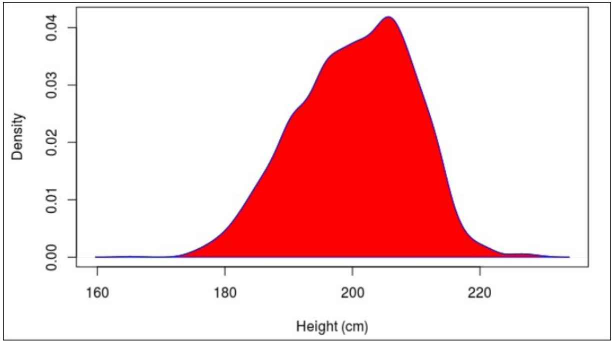
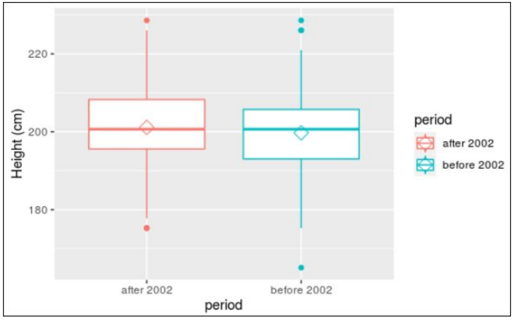


Figure 4: Density plot of Height before vs after 2002



As time goes by, I expect to see the younger generation dominate the older in height. I used the draft year 2002 to split players into 2 groups (each includes roughly 900 players) to compare their height distribution. The result boxplot shown in Figure 4 seems to confirm my speculation. The height mean of the younger group (201.1cm) is higher than that of the older (199.7cm). Overall, it can said that the height distribution of the younger group is more elevated than that of the older group, which implies that players drafted after 2002 are generally taller than their older counterpart.

Figure 5: Distribution of Weight (in pounds)

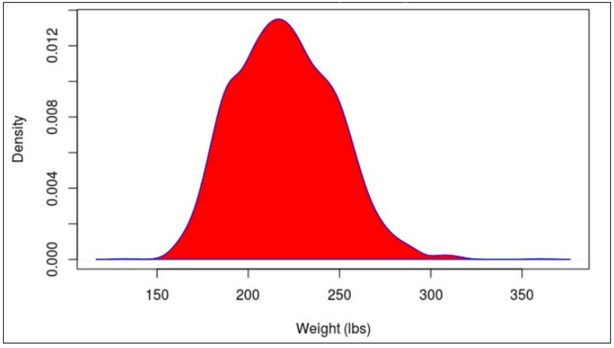
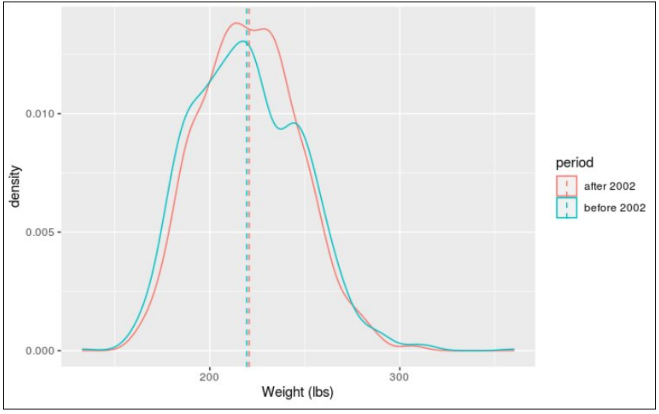


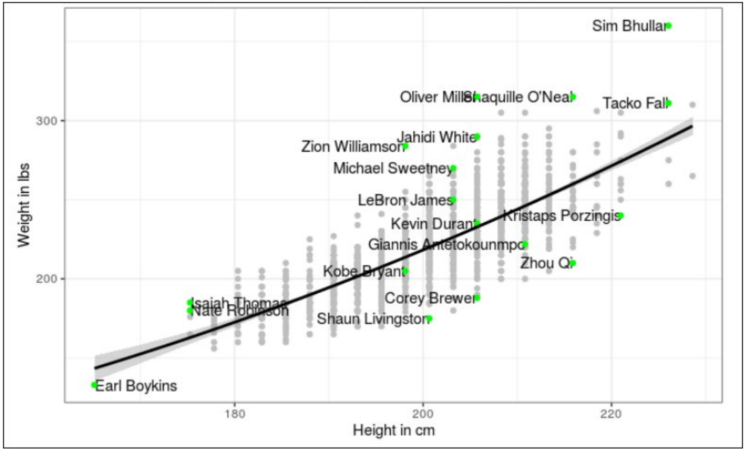
Figure 6: Density plot of Weight before vs after 2002



I address the same question for **Weight**, and the result is shown in Figure 5 and Figure 6. The most popular weight range is about 200-230 pounds, and the mean weight for two group are roughly similar. However, the distribution of weight for the younger group are more tightened to the middle, implying a higher percentage of the heavier players, compared to the players drafted before 2002. I use a density plot because the difference is not as recognizable in boxplot.

To conclude this section, I plotted to see how **Height** and **weight** are correlated, using a scatter plot in Figure 8. An apparent upward trend is seen in the relationship between **Height** and **Weight**. This chart also exposes the outliers in my dataset. There are players who are very heavy for their frame, namely Zion Williamson, Oliver Miller, Shaquille O'Neal, and Sim Bhullar. This can pose huge stress on their knees and ankles, leading to injuries. There are also players who are too skinny, such as Corey Brewer, Shaun Livingston, Zhou Qui and Kristaps Porzingis.

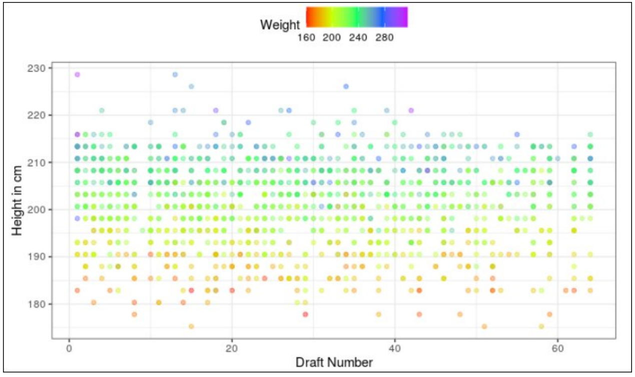
Figure 7: Height and Weight



4.2 Draft Number and Height, Weight

I attempt to find a relationship between **Draft Number** and **Height**. but there are no significant pattern found (see Figure 8 belows). The same can be said for the relationship between Weight and Height.s

Figure 8: Draft Number vs Height, Weight



4.3 Nationality and Physique

In 1908 entries in my dataset, there are only 377 foreign players (players who are not from the US). My theory is that foreign players must be physically dominant in order to be noticed by the NBA scouts.

Figure 9: Height of US vs non-US players

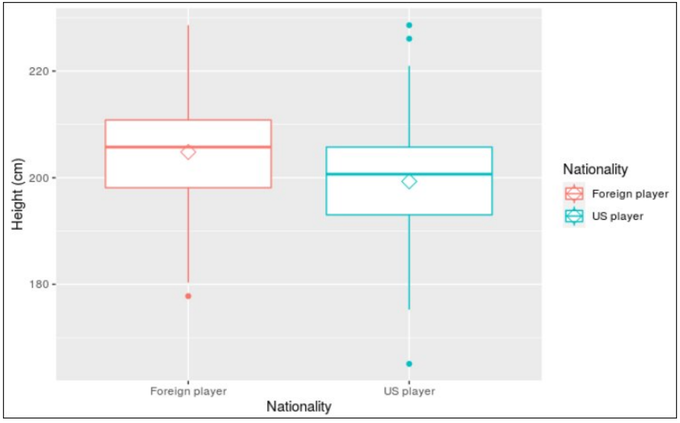
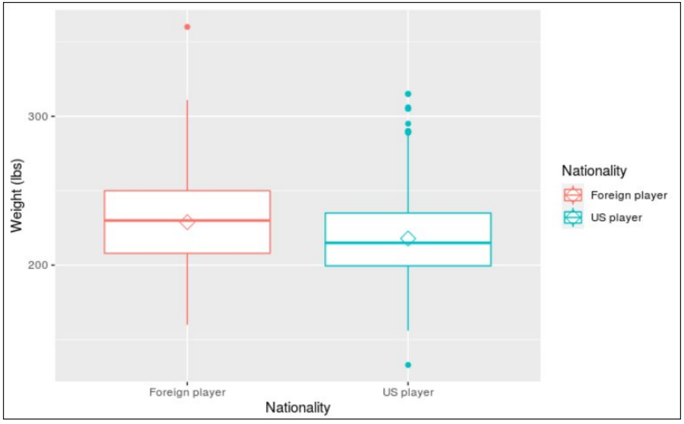


Figure 10: Weight of US vs non-US players



It is convincing from the density distribution plots in Figure 9 and 10 that non-US players are more dominant in height and weight: their distribution are more elevated than that of the US counterpart, and comparisons for mean height and weight also agree with my speculation. For example, according to figure 9, 50% of the foreign group is taller than 75% of the US group.

4.4 Physique and PER

In this section, I will use the PER data for the regular season 2018-2019 and 2005-2006 for comparison. I hope to see some differences between two seasons as the game is supposedly modernized now.

Figure 11: Height vs PER regular season 2018-2019

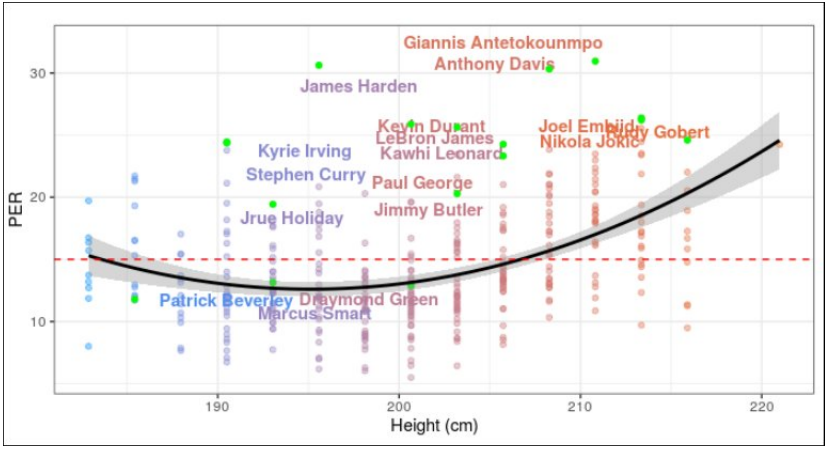
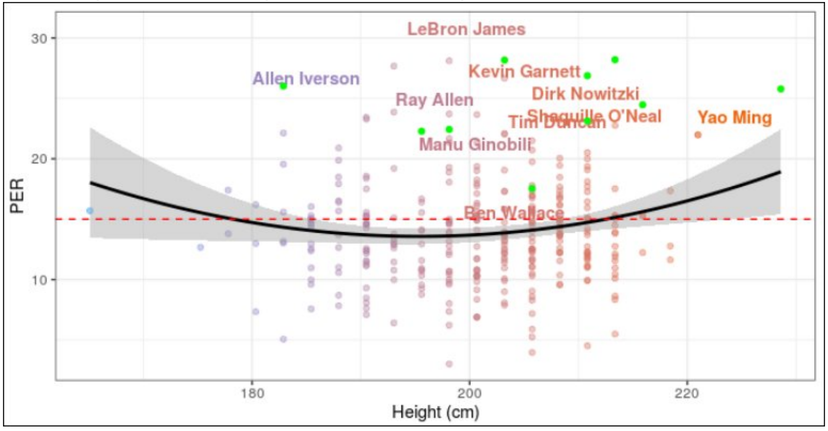


Figure 12: Height vs PER regular season 2005-2006



The result is shown in Figure 11 and 12. The color of the dots are simply based on **Height** for easier examination. Both charts reveals an interesting trend that bigger does not always mean better: it seems to be the opposite for players with height below 195cm. As the convex smoothing line implies, taller players (taller than 2m) seems to score higher PER. The trend is more recognizable in regular season 2018-2019, compared to the older time. The elite group also includes players shorter than 2m, such as Allen Iverson (season 2005-2006), Stephen Curry, James

Harden, and Kyrie Irving (season 2018-2019), whose skills are exceptional. But, in general, this group is dominated by big, tall, and strong forwards or centers (as the dominance of hot-colored dots implies).

Some of the most defensively successful players are underated by PER, such as Ben Wallace (best defensive player season 2005-2006), Patrick Beverley, Draymond Green, and Marcus Smart (season 2018-2019). The rating for Draymond Green is about 10 only, yet this is one of the best defensive players in the league in recent years.

5 Conclusion

In the end, it seems that heavier and taller players are becoming more dominant in the NBA, not only in quantity but also in quality: the highest achievers are usually the strongest players (power forwards or centers). Next, foreign ballers tend to be heavier and taller than US players. Finally, there is no significant proof that draft number is affected by height or weight.

6 Reflection

After this assignment, I became more efficient at scraping and cleaning data using Python. I also got to practice drawing plots using ggplot in Rstudio. I believe that the visualization aspect of my assignment is not too distracting but informative and coherent.

In subsection **4.4 Physique and PER**, I did not draw the similar plots for weight because the trend for weight and PER is not as clear as that of height and weight and the report is already too long. And, in order to illustrate the distribution, I had to pick between density plot and boxplot. It seems that sometimes the difference is exposed more in density plot (see Figure 6).

Overall, I am satisfied with the result, but I am open to criticism as there are always rooms for improvement.

References

Wikipedia contributors. (2020). *Player efficiency rating — Wikipedia, the free encyclopedia.* `https://en.wikipedia.org/w/index.php?title=Player_efficiency_rating&oldid=951282757`. ([Online; accessed 17-April-2020])