Assignment 1b

Tung X. Nguyen

April 19, 2020

Contents

1 Introduction **3**

2 Data Wrangling **3**
2.1 Data sources 4
2.1.1 NBA players regular season stats from Official NBA Statistics and Advanced Analytics 4
2.1.2 NBA Draft Combine Anthro 4
2.1.3 Player Efficiency Rating (PER) 4
2.2 Data Cleaning and Transformation 5

3 Data Checking **6**

4 Data Exploration **7**
4.1 The distribution of Height and Weight 7
4.2 Draft Number and Height, Weight 9
4.3 Nationality and Physique 10
4.4 Physique and PER 11

References **11**

List of Figures

1 PER reference guide 5
2 NBA players from 2000 to 2020 6
3 NBA Draft Combine Anthro from 2000 to 2020 6
4 Density plot of Height 7
5 Density plot of Height before vs after 2005 7
6 Distribution of Weight 8

7 Density plot of Weight before vs after 2005 8
8 Height and Weight 9
9 Draft Number vs Height, Weight 9
10 Height of US vs non-US players 10
11 Weight of US vs non-US players 10

1 Introduction

1

Basketball easily is one of the most favored kinds of sports ever been invented. The game is the most dominant in the United States, where the National Basketball Association, arguably the most competitive basketball league of the world, headquaters. Each regular season usually starts in October and finish in April. Then come the play-off, lasting from April to June and finally, the finals. The draft also happens in June, when young promising talents from over the world waiting for their name to be called.

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. Even though basketball IQ and tatics play a vital role, it is the muscle that delivers the hardwork. NBA basketballs are, in general, not only huge but also exceptionally fit. This was not always the case. In the past century, 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 (2019-2020) by position, by height, by weight, wing size ...?
- How physique can affect the draft number of a player? For instance: do bigger players get more chance of being picked in the first round?
- How physique and Player Efficiency Rating correlated?
- Is there any differences in the physique of non-US players to US players?

2 Data Wrangling

The data used for my projects are all scraped from online basketball websites mentioned in the subsection **Data sources**. I used Python with the help of libraries including Selenium, BeautifulSoup, pandas and csv to automatically surf the websites, scan for data, and finally write them down to dataframe that can be stored in text files and reused in the future.

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.

2.1.2 NBA Draft Combine Anthro

Link *stats.nba.com/draft/combine-anthro/*

This is a more detail table showing players' physicality index such as body fat percentage, wingspan, hands' length and width, height with shoes on and off. Not all players in the NBA show up in the Draft Combine Anthro. This data only applies to young talents who are going to be drafted for the same year or the year after. This data is used to complement the bio dataset above. It is important to accept that the wingspan, hands' length and width, and standing reach are less likely to change over seasons, but the body fat percentage is.

2.1.3 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.

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.

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

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).

For the NBA Draft Combine Anthro data, I also perform union to merge across seasons. It is worth noticing that the NBA Draft Combine Anthro started out with only Standing reach, Weight, and Wingpsan. Body fat were then introduced in the season 2003-2004. Finally, in the season 2010-2011, Hand length and Hand width were added to the measurement. For this dataset, I only keep data about wingspan, hand length and width, standing reach and body fat percentage. I also

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

Figure 3: NBA Draft Combine Anthro from 2000 to 2020

Player	BODY FAT %	\n HAND LENGTH (inches)\n	HAND WIDTH (inches)	STANDING REACH	WINGSPAN	Draft Year	Wingspan_cm	Standing_reach_cm
Coby White	4.30%	7.75	9.00	8' 1.5"	6' 5"	2019	195.580	247.65
Kris Wilkes	4.90%	8.50	9.50	8' 7"	6' 10.75"	2019	210.185	261.62
Grant Williams	5.40%	9.00	10.50	8' 8.5"	6' 9.75"	2019	207.645	265.43
Zion Williamson	-%	-	-	NaN	NaN	2019	NaN	NaN
Dylan Windler	4.60%	8.25	9.50	8' 8.5"	6' 10"	2019	208.280	265.43

convert the wingspan and standing reach from feet-inch to centimeters like I did for the regular season dataset. The last five entries in the resulting dataframe is shown in Figure 3.

There are 1350 entries in the Draft Combine Anthro datasets. However, not all players in this dataset are NBA players. Therefore, I perform a left join between the NBA regular season dataset with this dataset to complete the physique dataset.

In the similar manner, I retrieve the PER data. After scraping html data for multiple times, I became relatively efficient at this skill.

3 Data Checking

There are entries in my dataset that are problematic. They 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.

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 4. This distribution is skewed to the right, and the most popular height range is from about 195 to 210 centimeters. The tallest players are Yao Ming and Shawn Bradley (7f6 or 2.29m), while the shortest player is Earl Boykins (5f5 or 1.65m)

Figure 4: Density plot of Height

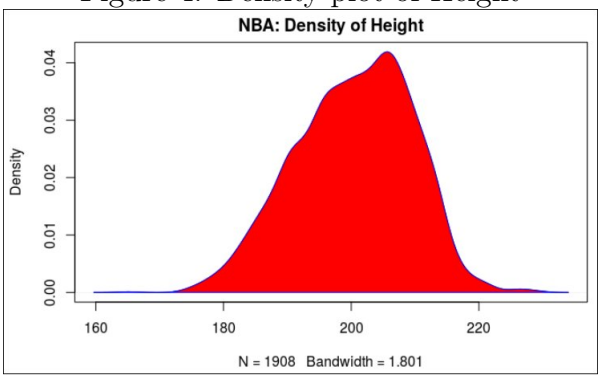
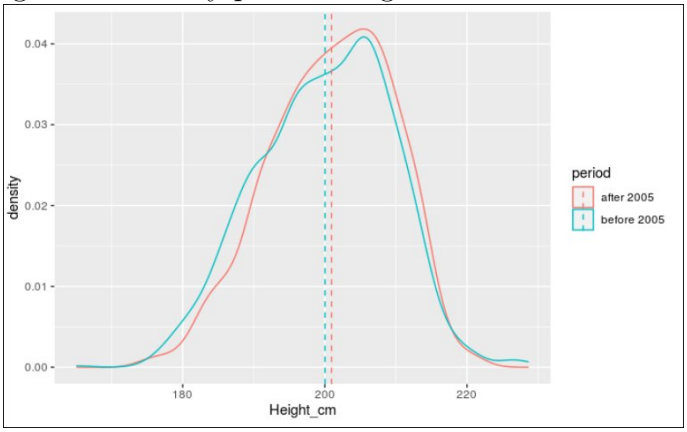


Figure 5: Density plot of Height before vs after 2005



As the basketball game progresses, I expect to see the younger generation dominate the older in height. I used the draft year 2005 to split players into 2 groups to compare their height distribution. The result, shown in Figure 5, seems to confirm my speculation. The height mean of the younger group (200.9cm) is

slightly higher than that of the older (200.2cm). The most height distribution of the young is also more tightened to the middle and shifted to the right, compared to the line for the older group, which implies that players drafted after 2005 are generally taller than their older counterpart.

Figure 6: Distribution of Weight

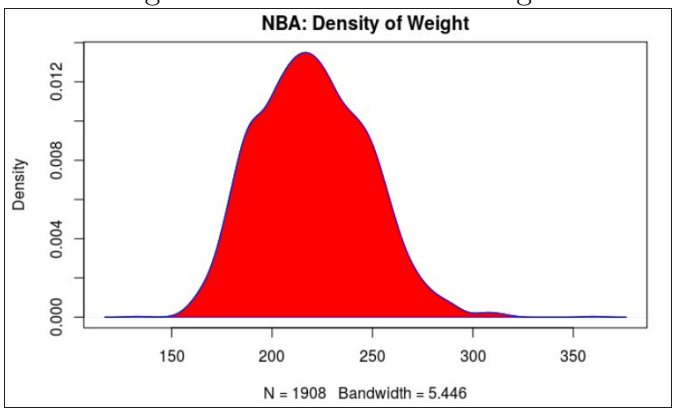
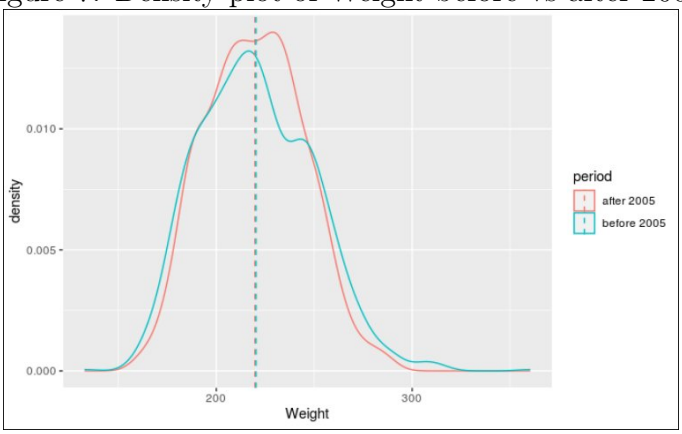


Figure 7: Density plot of Weight before vs after 2005

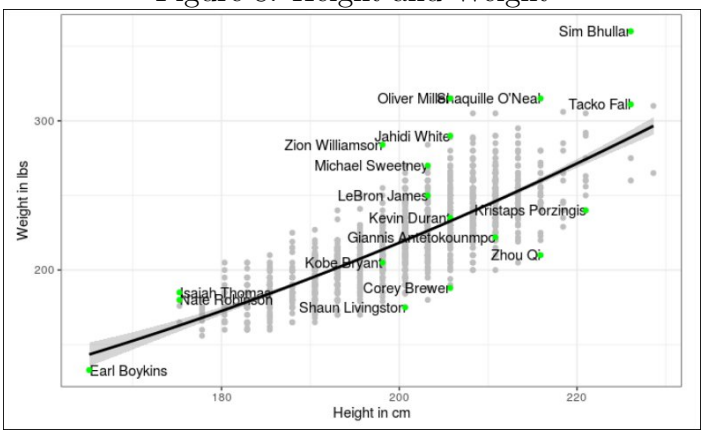


I address the same question for **Weight**, and the result is shown in Figure 6 and Figure 7. The most popular weight range is about 200-230 pounds, and the mean weight for two group are roughly similar. However, the distribution of weight for the younger group are more tightened to the middle, implying a higher percentage of the heavier players, compared to the players drafted before 2005.

To conclude this section, I plotted to see how **Height** and **weight** are correlated, using a scatter plot.

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

Figure 8: Height and Weight



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.

There is not enough wingspan data for all of the player in my dataset (the earliest draft combine data available is from 2000). Therefore I conclude this section and move to the next part.

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 9). The same can be said for the relationship between Weight and Height

Figure 9: 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 10: Height of US vs non-US players

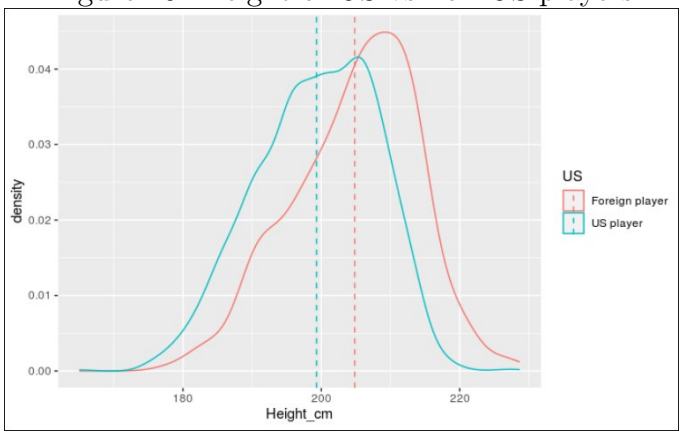
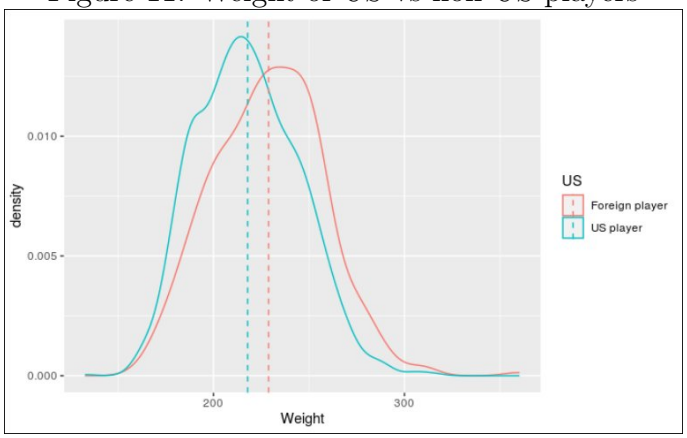


Figure 11: Weight of US vs non-US players



It is convincing from the density distribution plots that non-US players are more dominant in height and weight: their distribution are shifted to the right of the US counterpart, and comparisons for mean height and weight also agree with my speculation.

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.

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])