NBA Player Analysis 1996-2022

Jack Schofield

07-11-2022

Exploring the evolution of NBA Players

Introduction The National Basketball Association (NBA) is widely regarded as the premier basketball league in the world. Created in 1946, the league now consists of 30 teams from the United States and Canada, featuring the best players from around the globe. Now a truly global league, the last four players to win the Most Valuable Player (MVP) Award have hailed from outside the United States - an impressive achievement given that only 12% of MVPs were born outside of the United States.

NBA matches and pre-season matches have most recently taken in London, Paris, Abu Dhabi, Mexico, Japan and China, further cementing basketballs status as the third most watched sport in the world, behind football and cricket (source).

How has the game changed? With an increase of players entering the league displaying different cultural backgrounds to the game, and the increase in data analytics that has already led to a shift in playing still that has seen the traditional 'big' move out of the game, an increase in 3 point shooting and free throw attempts, just how has this affected those coming into the game? Whilst also looking at the draft profile of those entering the NBA and the value that higher picks give, specifically, I'll look to answer these questions:

1. How has the height profile changed for those entering the NBA via the draft?
2. Given the increase in higher value shots, how has the profile for points scoring changed?
3. How has the number of global superstars entering the NBA changed?

Ultimately, this deep dive should give a better profile to selecting the right player in the NBA Draft.

Getting started The data source for this set is 'NBA Players' by Justinas Cirtautas on Kaggle. In this instance, undrafted players won't be included in the final analysis but do appear in the dataset.

Initial R Code We'll start by getting all of our software setup and loaded.

```
#Install tidyverse
install.packages("tidyverse")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --

## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
```

```
## x dplyr::lag()    masks stats::lag()
```

```r
#Install and load ggplot
install.packages("ggplot2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
## also installing the dependencies 'lifecycle', 'vctrs'
```

```r
library(ggplot2)

#Load dataset
library(readr)
all_seasons <- read_csv("all_seasons.csv")
```

```
## New names:
## Rows: 12305 Columns: 22
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (8): player_name, team_abbreviation, college, country, draft_year, draf... dbl
## (14): ...1, age, player_height, player_weight, gp, pts, reb, ast, net_ra...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

Now that our data is all loaded up, let's start checking to make sure it's all consistent and that there are no errors in the data.

```r
#Check dataset
glimpse(all_seasons)
```

```
## Rows: 12,305
## Columns: 22
## $ ...1              <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15~
## $ player_name       <chr> "Dennis Rodman", "Dwayne Schintzius", "Earl Cureton"~
## $ team_abbreviation <chr> "CHI", "LAC", "TOR", "DAL", "MIA", "HOU", "LAL", "LA~
## $ age               <dbl> 36, 28, 39, 24, 34, 38, 25, 28, 29, 28, 27, 27, 28, ~
## $ player_height     <dbl> 198.12, 215.90, 205.74, 203.20, 205.74, 200.66, 198.~
## $ player_weight     <dbl> 99.79024, 117.93392, 95.25432, 100.69742, 108.86208,~
## $ college           <chr> "Southeastern Oklahoma State", "Florida", "Detroit M~
## $ country           <chr> "USA", "USA", "USA", "USA", "USA", "USA", "USA", "US~
## $ draft_year        <chr> "1986", "1990", "1979", "1995", "1985", "1981", "199~
## $ draft_round       <chr> "2", "1", "3", "1", "1", "2", "1", "1", "Undrafted",~
## $ draft_number      <chr> "27", "24", "58", "9", "10", "29", "10", "27", "Undr~
## $ gp                <dbl> 55, 15, 9, 64, 27, 52, 80, 77, 71, 82, 9, 1, 13, 3, ~
## $ pts               <dbl> 5.7, 2.3, 0.8, 3.7, 2.4, 8.2, 17.2, 14.9, 5.7, 6.9, ~
## $ reb               <dbl> 16.1, 1.5, 1.0, 2.3, 2.4, 2.7, 4.1, 8.0, 1.6, 1.5, 0~
## $ ast               <dbl> 3.1, 0.3, 0.4, 0.6, 0.2, 1.0, 3.4, 1.6, 1.3, 3.0, 1.~
## $ net_rating        <dbl> 16.1, 12.3, -2.1, -8.7, -11.2, 4.1, 4.1, 3.3, -0.3, ~
## $ oreb_pct          <dbl> 0.186, 0.078, 0.105, 0.060, 0.109, 0.034, 0.035, 0.0~
## $ dreb_pct          <dbl> 0.323, 0.151, 0.102, 0.149, 0.179, 0.126, 0.091, 0.1~
## $ usg_pct           <dbl> 0.100, 0.175, 0.103, 0.167, 0.127, 0.220, 0.209, 0.2~
## $ ts_pct            <dbl> 0.479, 0.430, 0.376, 0.399, 0.611, 0.541, 0.559, 0.5~
## $ ast_pct           <dbl> 0.113, 0.048, 0.148, 0.077, 0.040, 0.102, 0.149, 0.0~
## $ season            <chr> "1996-97", "1996-97", "1996-97", "1996-97", "1996-97~
```

```r
#Check for null values
colSums(is.na(all_seasons))
```

```
##               ...1       player_name team_abbreviation               age
##                  0                 0                 0                 0
##      player_height     player_weight           college           country
##                  0                 0                 5                 0
##         draft_year       draft_round      draft_number                gp
##                  0                 0                 0                 0
##                pts               reb               ast        net_rating
##                  0                 0                 0                 0
##          oreb_pct          dreb_pct           usg_pct            ts_pct
##                  0                 0                 0                 0
##            ast_pct            season
##                  0                 0
```

```r
#Use summary function to check high and low values for errors
summary(all_seasons)
```

```
##       ...1        player_name        team_abbreviation       age
##  Min.   :    0   Length:12305       Length:12305       Min.   :18.00
##  1st Qu.: 3076   Class :character   Class :character   1st Qu.:24.00
##  Median : 6152   Mode  :character   Mode  :character   Median :26.00
##  Mean   : 6152                                         Mean   :27.08
##  3rd Qu.: 9228                                         3rd Qu.:30.00
##  Max.   :12304                                         Max.   :44.00
##  player_height   player_weight      college            country
##  Min.   :160.0   Min.   : 60.33   Length:12305       Length:12305
##  1st Qu.:193.0   1st Qu.: 90.72   Class :character   Class :character
##  Median :200.7   Median : 99.79   Mode  :character   Mode  :character
##  Mean   :200.6   Mean   :100.37
##  3rd Qu.:208.3   3rd Qu.:108.86
##  Max.   :231.1   Max.   :163.29
##   draft_year        draft_round        draft_number             gp
##  Length:12305       Length:12305       Length:12305       Min.   : 1.00
##  Class :character   Class :character   Class :character   1st Qu.:31.00
##  Mode  :character   Mode  :character   Mode  :character   Median :57.00
##                                                           Mean   :51.29
##                                                           3rd Qu.:73.00
##                                                           Max.   :85.00
##       pts              reb              ast           net_rating
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   :-250.000
##  1st Qu.: 3.600   1st Qu.: 1.800   1st Qu.: 0.600   1st Qu.:  -6.400
##  Median : 6.700   Median : 3.000   Median : 1.200   Median :  -1.300
##  Mean   : 8.173   Mean   : 3.559   Mean   : 1.814   Mean   :  -2.256
##  3rd Qu.:11.500   3rd Qu.: 4.700   3rd Qu.: 2.400   3rd Qu.:   3.200
##  Max.   :36.100   Max.   :16.300   Max.   :11.700   Max.   : 300.000
##     oreb_pct          dreb_pct        usg_pct           ts_pct
##  Min.   :0.00000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.02100   1st Qu.:0.096   1st Qu.:0.1490   1st Qu.:0.4800
##  Median :0.04100   Median :0.131   Median :0.1810   Median :0.5240
##  Mean   :0.05447   Mean   :0.141   Mean   :0.1849   Mean   :0.5111
##  3rd Qu.:0.08400   3rd Qu.:0.180   3rd Qu.:0.2170   3rd Qu.:0.5610
##  Max.   :1.00000   Max.   :1.000   Max.   :1.0000   Max.   :1.5000
##     ast_pct          season
```

```
##  Min.   :0.0000   Length:12305
##  1st Qu.:0.0660   Class :character
##  Median :0.1030   Mode  :character
##  Mean   :0.1314
##  3rd Qu.:0.1780
##  Max.   :1.0000
```

```
#Check column headers
colnames(all_seasons)
```

```
##  [1] "...1"              "player_name"       "team_abbreviation"
##  [4] "age"               "player_height"     "player_weight"
##  [7] "college"           "country"           "draft_year"
## [10] "draft_round"       "draft_number"      "gp"
## [13] "pts"               "reb"               "ast"
## [16] "net_rating"        "oreb_pct"          "dreb_pct"
## [19] "usg_pct"           "ts_pct"            "ast_pct"
## [22] "season"
```

From the above, we can see that there are 5 null values within the college column, but seeing as we we won't be using that date, we can carry on by. If we did want to use it, we would just use the drop_na function. We can also see from the summary section that there are some values for True Scoring Percentage (ts_pct) that are above 1, an impossibility, so we will make sure to exclude those values when performing analyses.

Now, let's clean up our data and save this as a new dataframe.

```
#Rename height and weight column headers to include unit of measurement, add player height as m
all_seasons_clean <- rename(all_seasons, player_height_cm=player_height, player_weight_kg=player_weight)
  mutate(all_seasons, player_height_m=player_height_cm/100) %>%

  #Filter out results that exceed 1 (maximum possible as a true scoring percentage)
  filter(ts_pct<=1) %>%

  #Filter out undrafted players and errors in draft number (highest is 60)
  filter(draft_number %in% 1:60)
```

Great, now that we have that all saved, let's start digging down deeper into the data! Let's start with looking at question 1.

```
#Group by draft year to measure average height across those drafts (filtered from 1996 to ensure all dr
  filter(all_seasons_clean, draft_year >=1996) %>%
  group_by(draft_year) %>%
  summarize (mean_height_m = mean(player_height_m)) %>%
  print(n=26)
```

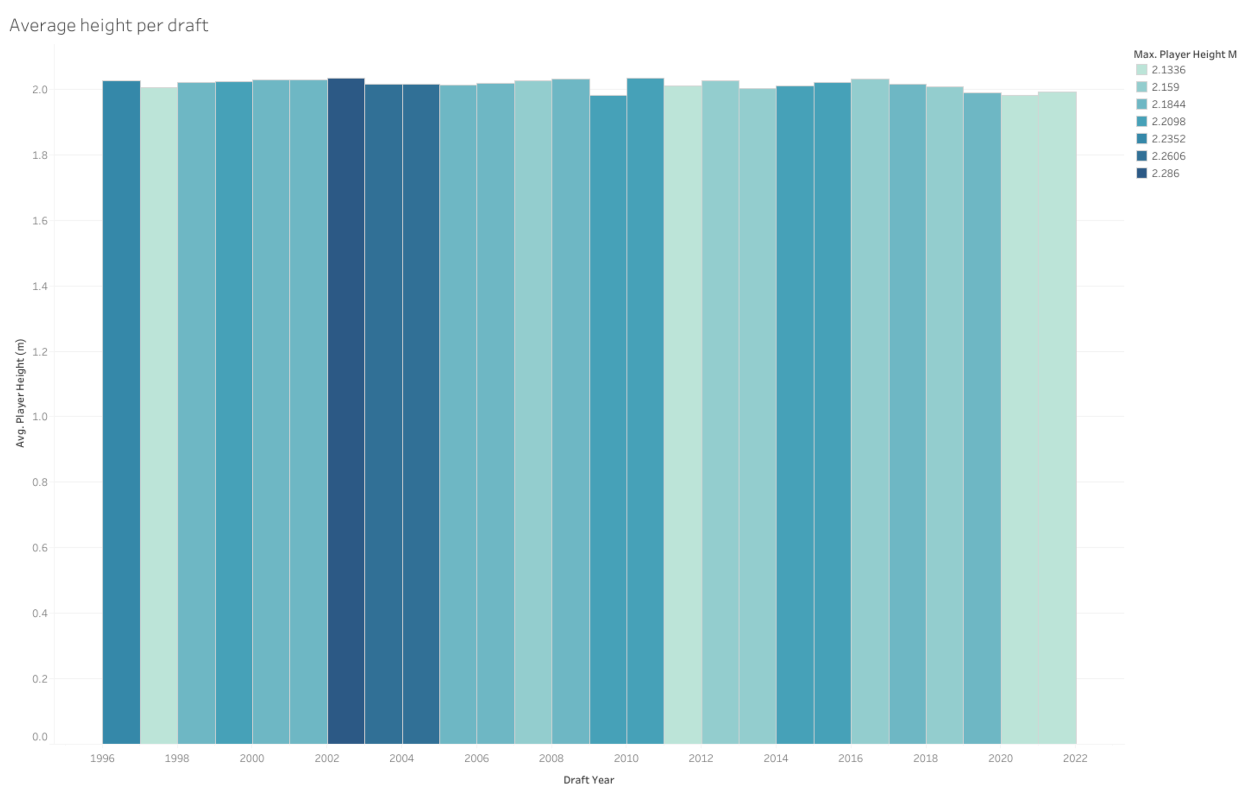
How has the height profile changed for those entering the NBA via the draft?

```
## # A tibble: 26 x 2
##    draft_year mean_height_m
##    <chr>              <dbl>
##  1 1996                2.01
##  2 1997                2.01
##  3 1998                2.02
##  4 1999                2.01
##  5 2000                2.02
##  6 2001                2.04
```

```
##  7 2002                      2.04
##  8 2003                      2.01
##  9 2004                      2.00
## 10 2005                      2.00
## 11 2006                      2.00
## 12 2007                      2.03
## 13 2008                      2.03
## 14 2009                      1.97
## 15 2010                      2.03
## 16 2011                      2.00
## 17 2012                      2.02
## 18 2013                      2.01
## 19 2014                      2.01
## 20 2015                      2.01
## 21 2016                      2.02
## 22 2017                      2.00
## 23 2018                      1.99
## 24 2019                      1.98
## 25 2020                      1.98
## 26 2021                      1.99
```

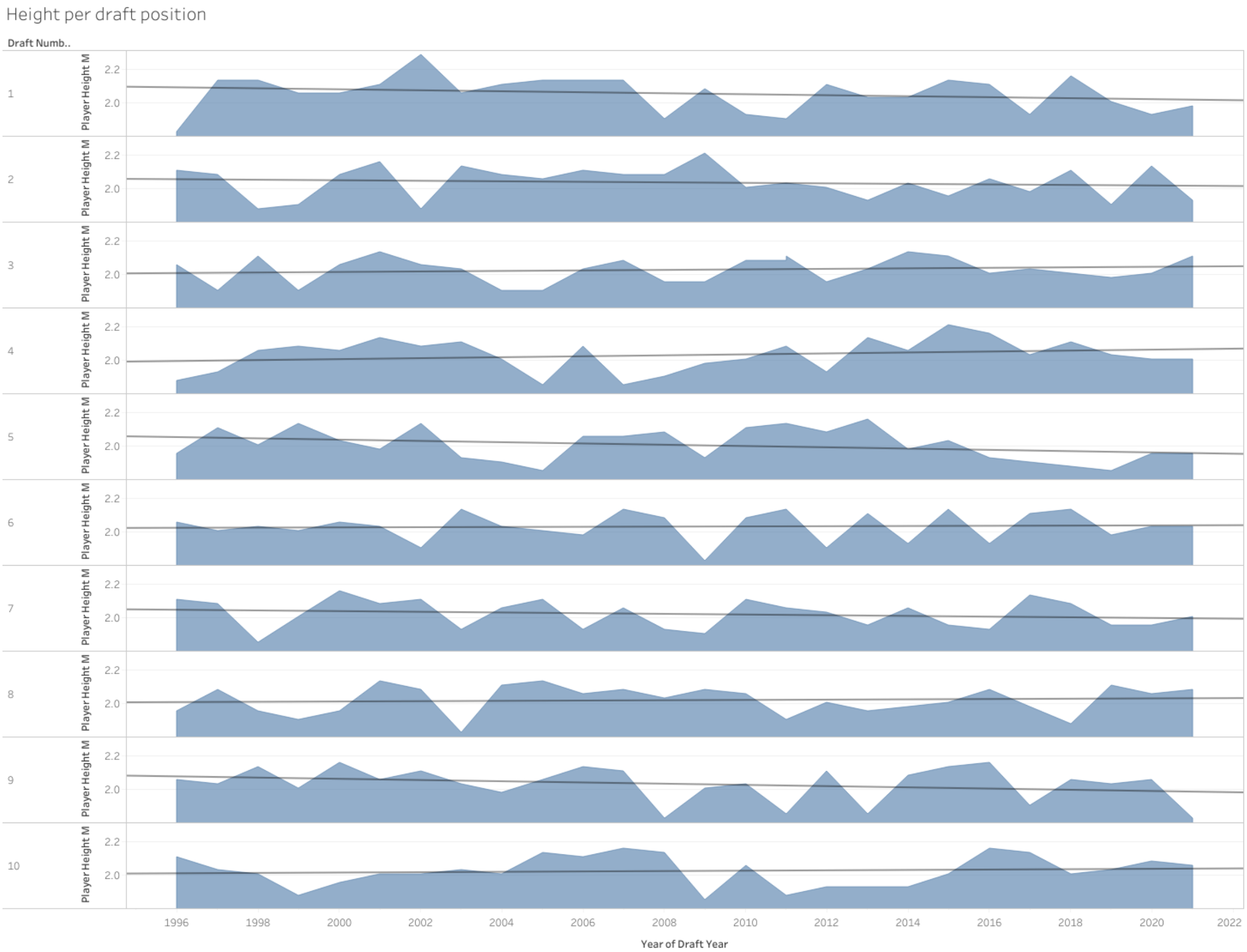
Whilst there's not an enormous reduction in height, the draft choices in the NBA are getting smaller. No draft has been as small as 2009, but the last 4 years have all clocked in at under 2m per player - directly correlating with the success of the Golden State Warriors dynasty and their free flowing movement game. Big centres tend to be less nimble, meaning that they can be easily isolated and targeted by 'small ball' line ups.

The visualisation below shows how little the height has varied, although the height of the tallest player has shown a marked decrease over the last decade.



The graph below indicates that there has started to be a slight drop off in height across most of the top

10 draft picks since 1996, although not all. This suggests that although the average height of draftees isn't changing rapidly, the value in which tall players have is diminishing.



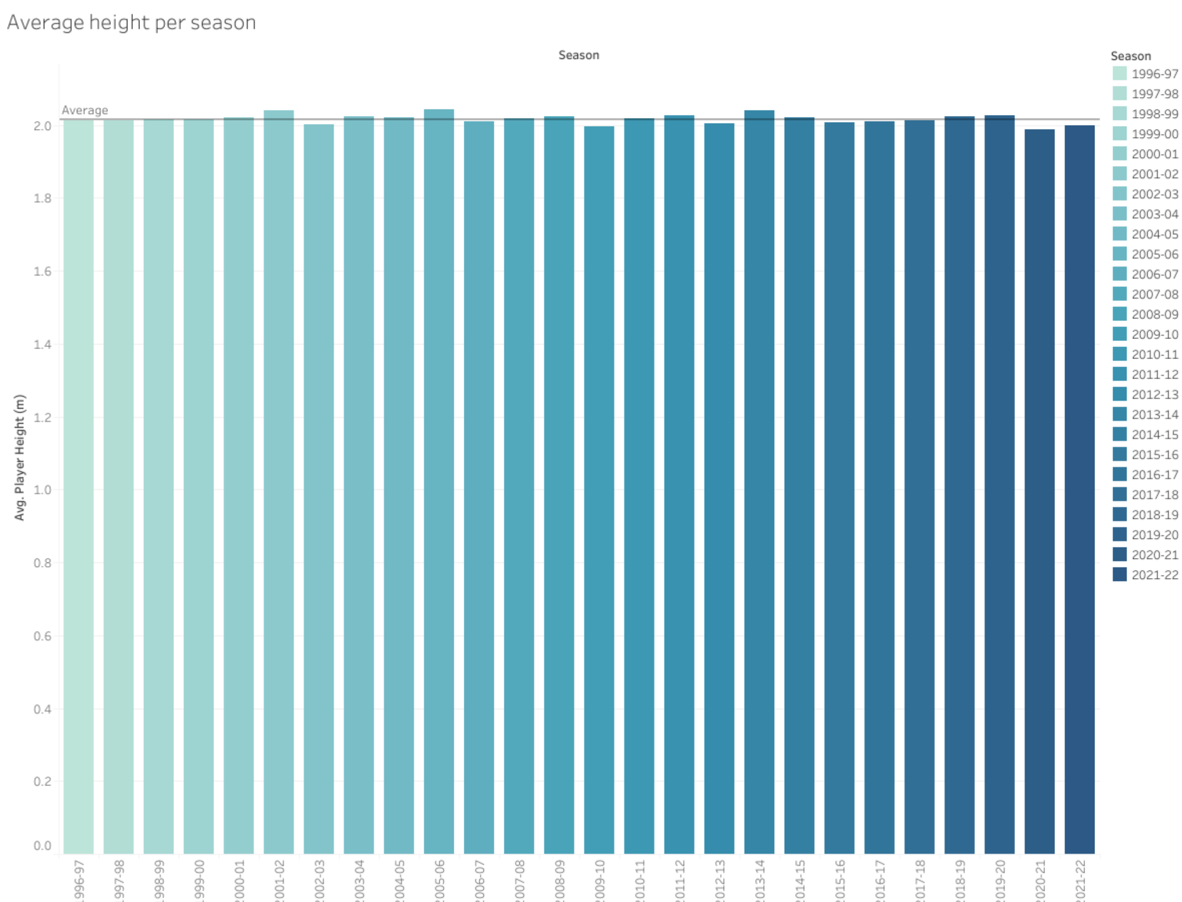

```r
#Group by season to measure average height
group_by(all_seasons_clean, season) %>%
summarize (mean_height_m = mean(player_height_m)) %>%
print(n=26)
```

```
## # A tibble: 26 x 2
##    season  mean_height_m
##    <chr>           <dbl>
##  1 1996-97          2.01
##  2 1997-98          2.01
##  3 1998-99          2.01
##  4 1999-00          2.02
##  5 2000-01          2.01
##  6 2001-02          2.02
##  7 2002-03          2.02
##  8 2003-04          2.02
##  9 2004-05          2.02
## 10 2005-06          2.02
## 11 2006-07          2.01
## 12 2007-08          2.02
## 13 2008-09          2.02
```

```
## 14 2009-10            2.01
## 15 2010-11            2.02
## 16 2011-12            2.01
## 17 2012-13            2.01
## 18 2013-14            2.01
## 19 2014-15            2.01
## 20 2015-16            2.01
## 21 2016-17            2.02
## 22 2017-18            2.01
## 23 2018-19            2.01
## 24 2019-20            2.00
## 25 2020-21            1.99
## 26 2021-22            1.99
```

Again, very little variation in the numbers but the last two NBA seasons have seen average height drop below 2m for the first time since at least 1996 (and, more than likely, ever).

Interestingly, on the below graph, I wasn't able to add a trend line, suggesting there are no discernible differences.



Clearly, the height of players being selected, and entering, the draft is reducing albeit at a slow rate. There will always be the need in some matches and match ups for a true big, it may just be that the skill-set of the traditional big is evolving. Nikola Jokic, the winner of the past two MVP Awards is 2.11m and has the passing ability of traditional point guard. His contemporary, Joel Embiid, is 2.13m tall and has such power and agility that he is rarely left exposed. Even as we look at the 2023 NBA Draft, the expected number 1 pick, Victor Wembanyama, is 2.19m tall and is a true 'unicorn'.

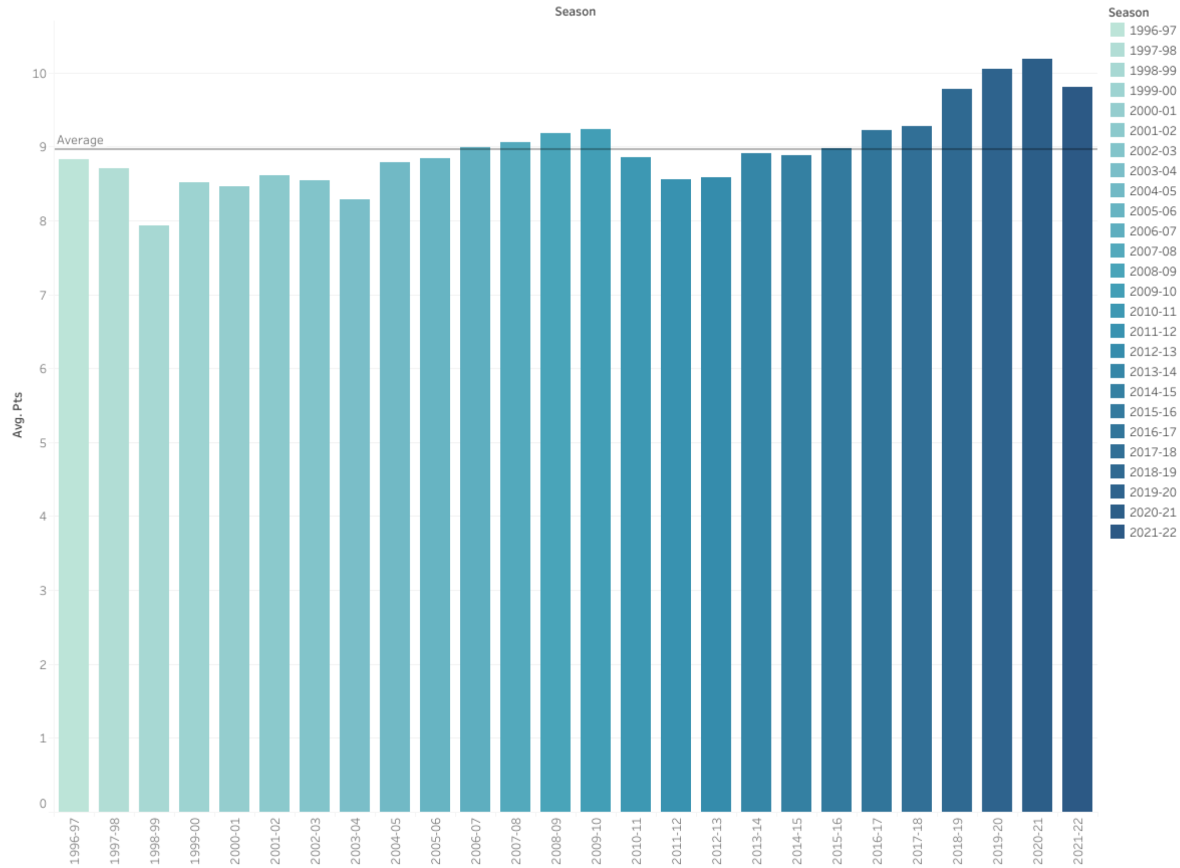
Is the NBA getting smaller? Yes, slightly, but more so that the value of taller players is diminishing with the exceptions of those who display modern NBA skills, such as the ability to space the floor.

Given the increase in higher value shots, how has the profile for points scoring changed? The arrival of Daryl Morey at the Houston Rockets changed the way the franchise played the game. A prolific believer in statistics and the part that 'Moneyball' played in changing the use of data in baseball, Daryl was the first statistics led General Manager (GM) in the NBA. Key to his philosophy as GM was increasing the number of high value shots - namely the three pointer and free throws. His data showed that mid-range and long-range pull ups (worth two points) were high risk and low reward, whereas the three pointer was high risk and high reward, with free throws low risk and high reward. But, has this changed the points profile of teams in the league?

```
#Group by season to measure points
group_by(all_seasons_clean, season) %>%
  summarize (mean_points_per_player = mean(pts)) %>%
  print(n=26)
```

```
## # A tibble: 26 x 2
##    season  mean_points_per_player
##    <chr>                    <dbl>
##  1 1996-97                   8.83
##  2 1997-98                   8.71
##  3 1998-99                   7.93
##  4 1999-00                   8.51
##  5 2000-01                   8.47
##  6 2001-02                   8.62
##  7 2002-03                   8.54
##  8 2003-04                   8.29
##  9 2004-05                   8.78
## 10 2005-06                   8.84
## 11 2006-07                   8.99
## 12 2007-08                   9.05
## 13 2008-09                   9.18
## 14 2009-10                   9.23
## 15 2010-11                   8.85
## 16 2011-12                   8.55
## 17 2012-13                   8.58
## 18 2013-14                   8.91
## 19 2014-15                   8.89
## 20 2015-16                   8.97
## 21 2016-17                   9.21
## 22 2017-18                   9.27
## 23 2018-19                   9.78
## 24 2019-20                  10.1
## 25 2020-21                  10.2
## 26 2021-22                   9.80
```

There is a reasonable amount of noise within the data. It's certainly true that the scoring rate has increased over the last few seasons, but this was following a slight decline from the late 00s. We can attribute the high scoring rate in 2019-20 to the boost in scoring during the bubble, and the following year to erratic attendances due to rolling lockdowns and restrictions. It is true that the last four seasons have been the highest scoring, and therefore it can be summised that scoring has increased and is likely to continue to do so.

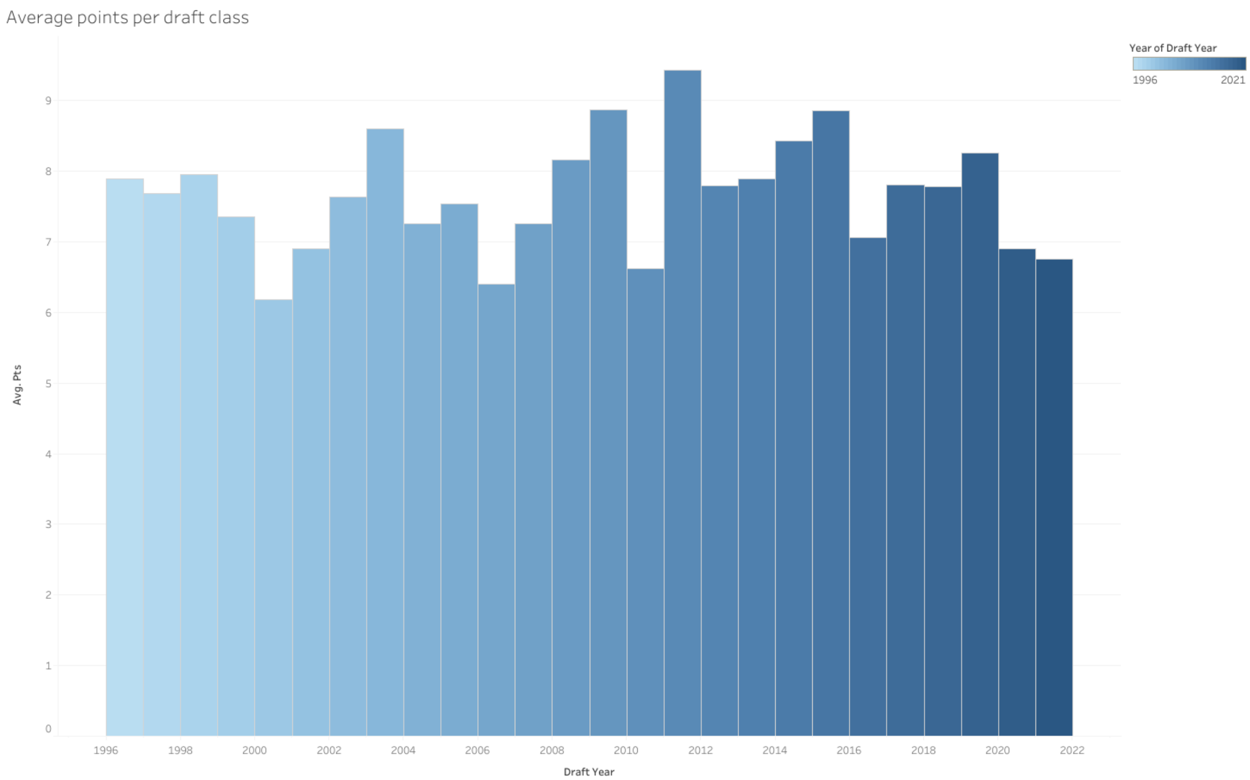



```
#Group by draft year to measure points
  filter(all_seasons_clean, draft_year >=1996) %>%
    group_by(draft_year) %>%
    summarize (mean_points_per_player = mean(pts)) %>%
    print(n=26)
```

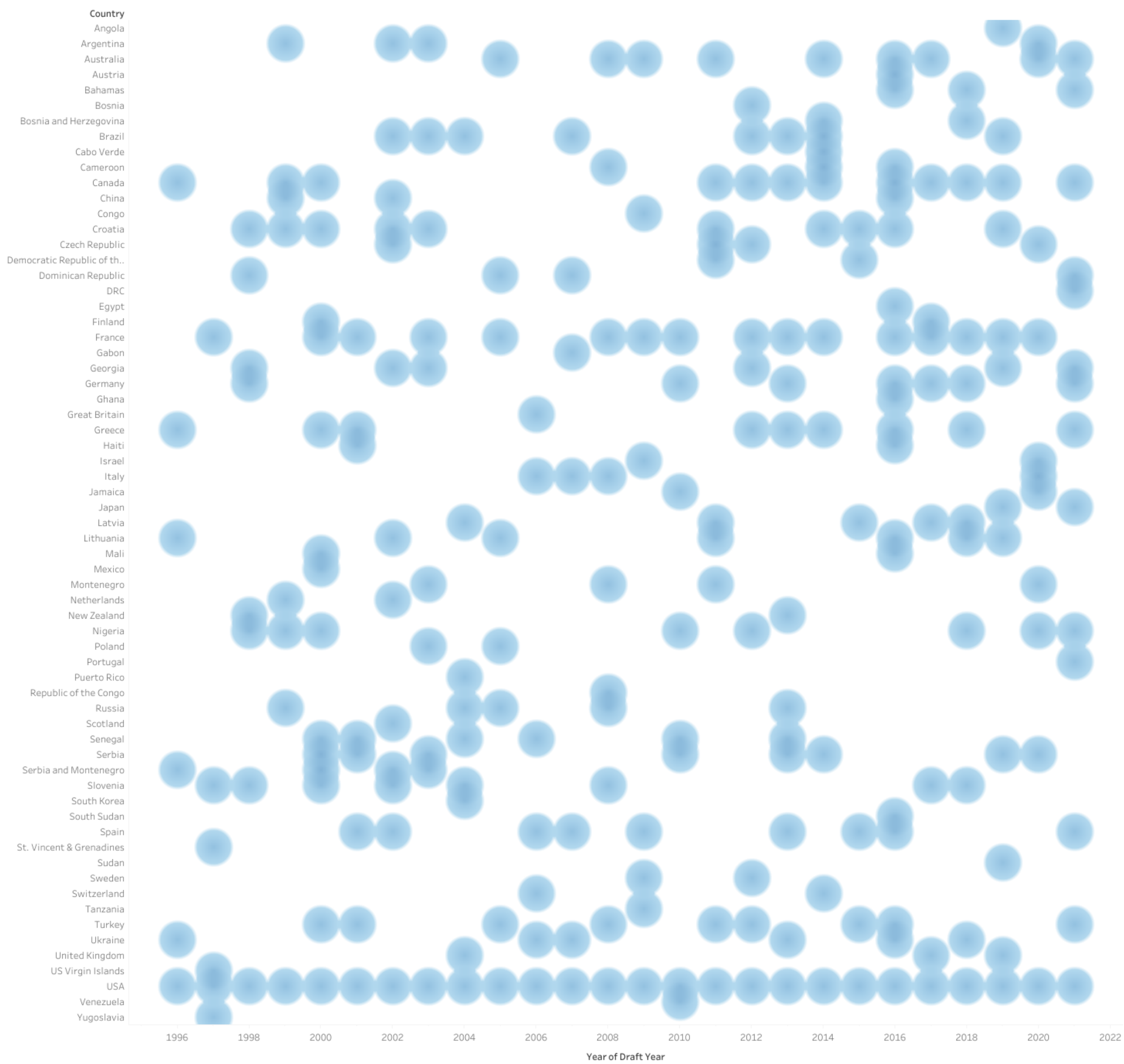
```
## # A tibble: 26 x 2
##    draft_year mean_points_per_player
##    <chr>                       <dbl>
##  1 1996                        10.3
##  2 1997                         8.73
##  3 1998                         9.37
##  4 1999                         9.61
##  5 2000                         7.03
##  6 2001                         8.63
##  7 2002                         8.44
##  8 2003                         9.84
##  9 2004                         9.01
## 10 2005                         9.03
## 11 2006                         8.13
## 12 2007                         9.45
## 13 2008                        10.1
## 14 2009                        10.1
## 15 2010                         8.26
## 16 2011                        10.5
## 17 2012                         9.24
```

```
## 18 2013                          8.46
## 19 2014                          9.71
## 20 2015                          9.32
## 21 2016                          8.00
## 22 2017                          8.52
## 23 2018                          8.87
## 24 2019                          7.78
## 25 2020                          6.67
## 26 2021                          6.75
```

Looking at the draft data, it would immediately suggest that scoring has been decreasing per draft class. However, as with any professional sports league, it takes time for a rookie to establish themselves in the side and begin to output on a consistent basis, so this data can be disregarded for now.



How has the number of global superstars entering the NBA changed? As can be seen from the visualisation below, the NBA is becoming an increasingly global game, with greater numbers from around the world entering into, and being selected from, the NBA Draft.



To see a full visual and filterable breakdown of country data, click here.

Conclusions To properly conclude this analysis, it would be really useful to remind ourselves of the questions that we set out to answer:

1. How has the height profile changed for those entering the NBA via the draft?
2. Given the increase in higher value shots, how has the profile for points scoring changed?
3. How has the number of global superstars entering the NBA changed?

Throughout the data we have seen that the height of the NBA isn't changing dramatically, but is very slowly starting to reduce. What we have seen is that the height of the top 10 picks in the NBA is starting to decrease (although with a few notable exceptions, like James Wiseman at number 2 in 2020), and therefore so is the value that NBA teams place on bigs with traditional skills. There will always be a place for a unicorn, with teams often scrambling to acquire and, sometimes, overpay them. In general, the big man is becoming less common in today's NBA.

With it, and the increase of smaller, more agile players, the average points per player is increasing. More work is needed to be done to determine if this is simply due to an increased number of turnovers, increased three point shooting, an increasing ability to get to the free throw line, an increase in pace of play, or something else. The NBA scoring rate is on the up, but is it sustainable?

Finally, the number of global players entering the NBA is rapidly increasing. As each new generation of players enters from abroad, the game grows globally and, coupled with the success of Jokic, Embiid and Giannis Antetokounmpo, skills worldwide are growing. National teams are growing and threatening the dominance of Team USA and this trend can be expected to continue. Further analysis could show the performance differences between US and non US players.

To summise, it is likely that NBA teams will continue to place greater emphasis on smaller players, reserving lower draft picks to select a traditional big man for the increasingly fewer times when they are needed. With this, the scoring rate will continue to improve, and teams will continue to place a greater importance on global scouting networks.

Future research into players We've already seen that the NBA is placing lesser value in taller players, traditional bigs, and that the average points per player is steadily climbing. Outside of this data, we know that teams are placing more emphasis on, and paying more money to, 3 & D wings - multi-faceted players who can score the value shots and defend well. The Toronto Raptors are steadily building themselves a team full of players that are 6'9", eager to create the next evolution of playing style in the NBA. For future research, it would be interesting and worthwhile to delve deeper into individual statistics such as three point percentage and individual defensive metrics, along with how the league has changed how it pays different types of player. It would also be really interesting to analyse the make up of championship teams, where they draft and sign players from, and the key metrics that create that winning DNA.