A statistical justification for tanking

Kuanyu Lai

2023-04-30

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.1     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.2     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors
```

```
library(networkD3)
library(patchwork)
library(dplyr)
```

```
draft2 <- read.csv('statisticsplaybook-main/statisticsplaybook-main/draft2.csv',header = TRUE, stringsAsFactors = FALSE)
```

```
glimpse(draft2)
```

```
## Rows: 291
## Columns: 18
## $ Rk       <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18…
## $ Year     <int> 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2…
## $ Pk       <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18…
## $ Tm       <chr> "LAC", "MEM", "OKC", "SAC", "MIN", "MIN", "GSW", "NYK", "TOR"…
## $ Player   <chr> "Blake Griffin", "Hasheem Thabeet", "James Harden", "Tyreke E…
## $ Age      <dbl> 20.106, 22.135, 19.308, 19.284, 18.252, 20.144, 21.108, 21.33…
## $ Pos      <chr> "F", "C", "G", "G-F", "G", "G", "G", "C-F", "G-F", "G", "F", …
## $ Born     <chr> "us", "tz", "us", "us", "es", "us", "us", "us", "us", "us", "…
## $ College  <chr> "Oklahoma", "UConn", "Arizona State", "Memphis", "0", "Syracu…
## $ From     <int> 2011, 2010, 2010, 2010, 2012, 2010, 2010, 2010, 2010, 2010, 2…
## $ To       <int> 2020, 2014, 2020, 2019, 2020, 2012, 2020, 2017, 2020, 2018, 2…
## $ G        <int> 622, 224, 826, 594, 555, 163, 699, 409, 813, 555, 153, 535, 4…
## $ MP       <dbl> 34.8, 10.5, 34.3, 30.7, 30.9, 22.9, 34.3, 18.8, 34.2, 30.3, 1…
## $ WS       <dbl> 75.2, 4.8, 133.3, 28.4, 36.4, -1.1, 103.2, 16.4, 66.5, 29.9, …
## $ WS.48    <dbl> 0.167, 0.099, 0.226, 0.075, 0.102, -0.015, 0.207, 0.102, 0.11…
## $ Born2    <chr> "USA", "World", "USA", "USA", "World", "USA", "USA", "USA", "…
## $ College2 <int> 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ Pos2     <chr> "Forward", "Center", "Guard", "Swingman", "Guard", "Guard", "…
```

Create new variable

The derived variable Pk2 equals 1-5 when the original variable equals any number between 1 and 5. The %in% operator in R identifies if an element, such as a number, is included in a vector or data frame. If affirmative, Pk2 is assigned the value 1-5; if otherwise, the next line of code is read, and so forth. Pk2 equals 6-10 if Pk equals any number between 6 and 10. Pk2 equals 11-15 if Pk equals any number between 11 and 15. Pk2 equals 16-20 if Pk equals any number between 16 and 20. Pk2 equals 21-25 if Pk equals any number between 21 and 25. Pk2 equals 26-30 if Pk equals any number between 26 and 30. If the original variable Pk equals anything other than a number between 1 and 30, the new variable Pk2 will equal NA.

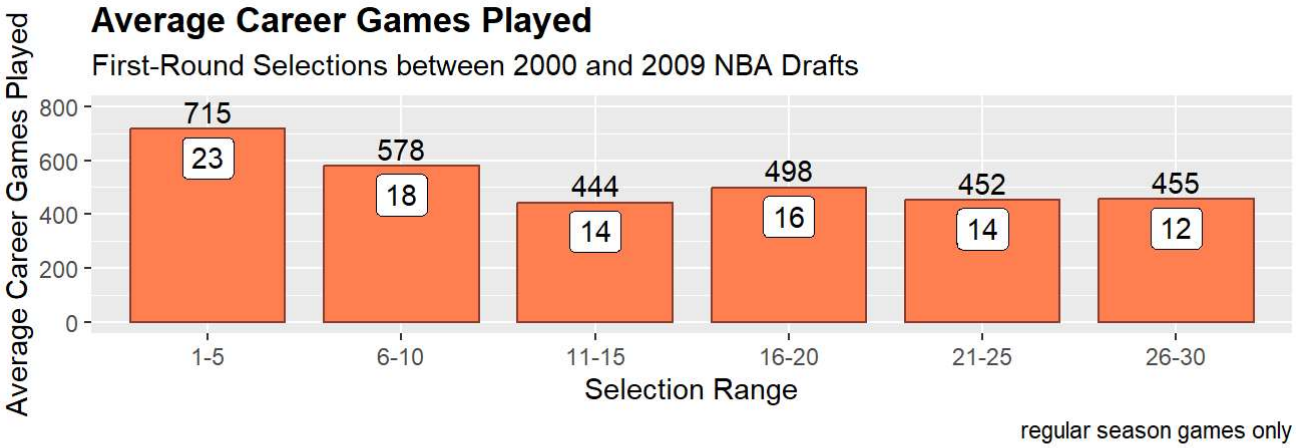
```
mutate(draft2, Pk2 = ifelse(Pk %in% 1:5, "1-5",
                     ifelse(Pk %in% 6:10, "6-10",
                     ifelse(Pk %in% 11:15, "11-15",
                     ifelse(Pk %in% 16:20, "16-20",
                     ifelse(Pk %in% 21:25, "21-25",
                     ifelse(Pk %in% 26:30, "26-30", "NA")))))) -> draft2
draft2$Pk2 <- as.factor(draft2$Pk2)
```

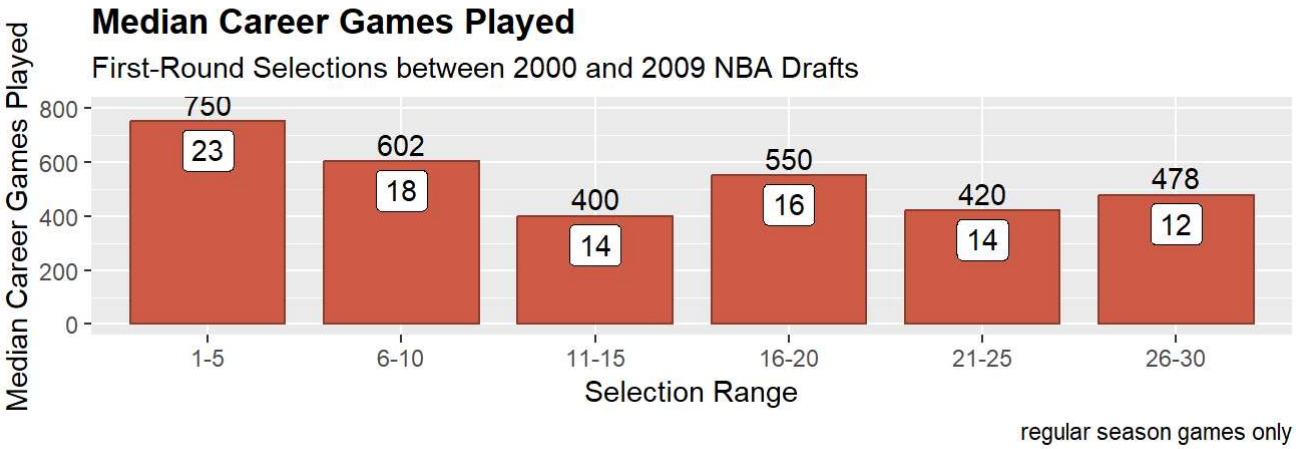
```
#table 1
sumG <- sum(draft2$G)
table1 <- draft2 %>%
  dplyr::group_by(Pk2) %>%
  dplyr::summarize(mean = mean(G),
                   median = median(G),
                   pct = sum(G)/sumG)
table1
```

Pk2 <fct>	mean <dbl>	median <dbl>	pct <dbl>
1-5	715.9000	750.5	0.2336641
11-15	444.0408	400.0	0.1420328
16-20	498.0400	550.0	0.1625563
21-25	452.8200	420.0	0.1477969
26-30	455.8333	478.0	0.1249755
6-10	578.9800	602.5	0.1889745

6 rows

```
g1 <- ggplot(table1, aes(x = Pk2, y = mean)) +
  geom_bar(stat = "identity", width = .8, fill = "coral", color = "coral4") +
  labs(title = "Average Career Games Played",
       subtitle = "First-Round Selections between 2000 and 2009 NBA Drafts",
       x = "Selection Range", y = "Average Career Games Played",
       caption = "regular season games only") +
    scale_x_discrete(limits = c("1-5", "6-10", "11-15", "16-20", "21-25", "26-30"),
                     labels = c("1-5", "6-10", "11-15", "16-20", "21-25", "26-30")) +
  geom_text(aes(label = trunc(mean), vjust = -0.3)) +
  geom_label(aes(label = trunc(pct*100), vjust = 1.2)) +
  ylim(0, 800) +
  theme(plot.title = element_text(face = "bold"))

g2 <- ggplot(table1, aes(x = Pk2, y = median)) +
  geom_bar(stat = "identity", width = .8, fill = "coral3", color = "coral4") +
  labs(title = "Median Career Games Played",
       subtitle = "First-Round Selections between 2000 and 2009 NBA Drafts",
       x = "Selection Range", y = "Median Career Games Played",
       caption = "regular season games only") +
    scale_x_discrete(limits = c("1-5", "6-10", "11-15", "16-20", "21-25", "26-30"),
                     labels = c("1-5", "6-10", "11-15", "16-20", "21-25", "26-30")) +
  geom_text(aes(label = trunc(median), vjust = -0.3)) +
  geom_label(aes(label = trunc(pct*100), vjust = 1.2)) +
  ylim(0, 800) +
  theme(plot.title = element_text(face = "bold"))
g1 + g2 + plot_layout(ncol = 1)
```





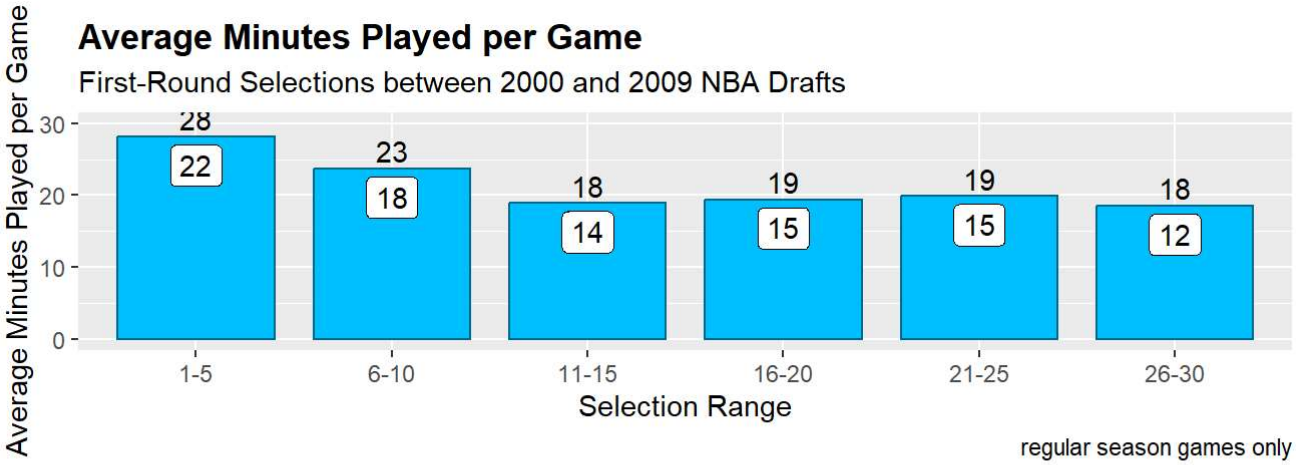
Key findings:

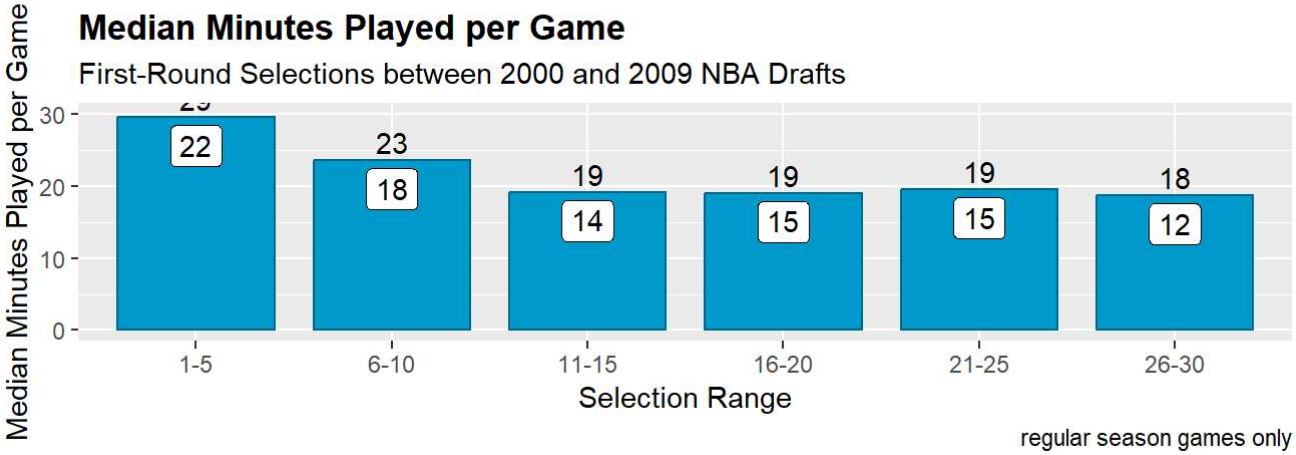
1.Players selected first through fifth played in more regular season games, based on means and medians, than any other group of first-round selections; in fact, though these players represent about 17% of the draft2 record count, they collectively played in more than 23% of the regular season games.

2.Players then selected in spots 6 through 10 played in more regular season games than players selected later in the first round; they, too, represent approximately 17% of the draft2 record count, but no less than 18% of the total regular season games played. Therefore, about 34% of the first-round selections in draft2 account for over 41% of the total regular season games played.

3.In no other selection range does the percentage of regular season games played exceed or even equal their respective percentage of draft2 record counts; in other words, selection ranges 11-15, 16-20, 21-25, and 26-30 each contribute approximately 17% of the records to the draft2 data set, yet the percentage of total regular season games played across those four selection ranges is consistently less than 17%.

```
#minute play per game
sumMP <- sum(draft2$MP)
draft2 %>%
  dplyr::group_by(Pk2) %>%
  dplyr::summarize(mean = mean(MP),
            median = median(MP),
            pct = sum(MP)/sumMP) -> table2

mp1 <- ggplot(table2, aes(x = Pk2, y = mean)) +
  geom_bar(stat = "identity", width = .8, fill = "deepskyblue", color = "deepskyblue4") +
  labs(title = "Average Minutes Played per Game",
       subtitle = "First-Round Selections between 2000 and 2009 NBA Drafts",
       x = "Selection Range", y = "Average Minutes Played per Game",
       caption = "regular season games only") +
    scale_x_discrete(limits = c("1-5", "6-10", "11-15", "16-20", "21-25", "26-30"),
                     labels = c("1-5", "6-10", "11-15", "16-20", "21-25", "26-30")) +
  geom_text(aes(label = trunc(mean), vjust = -0.3)) +
  geom_label(aes(label = trunc(pct*100), vjust = 1.2)) +
  ylim(0, 30) +
  theme(plot.title = element_text(face = "bold"))

mp2 <- ggplot(table2, aes(x = Pk2, y = median)) +
  geom_bar(stat = "identity", width = .8, fill = "deepskyblue3", color = "deepskyblue4") +
  labs(title = "Median Minutes Played per Game",
       subtitle = "First-Round Selections between 2000 and 2009 NBA Drafts",
       x = "Selection Range", y = "Median Minutes Played per Game",
       caption = "regular season games only") +
    scale_x_discrete(limits = c("1-5", "6-10", "11-15", "16-20", "21-25", "26-30"),
                     labels = c("1-5", "6-10", "11-15", "16-20", "21-25", "26-30")) +
  geom_text(aes(label = trunc(median), vjust = -0.3)) +
  geom_label(aes(label = trunc(pct*100), vjust = 1.2)) +
  ylim(0, 30) +
  theme(plot.title = element_text(face = "bold"))

mp1 + mp2 + plot_layout(ncol = 1)
```





Key findings:

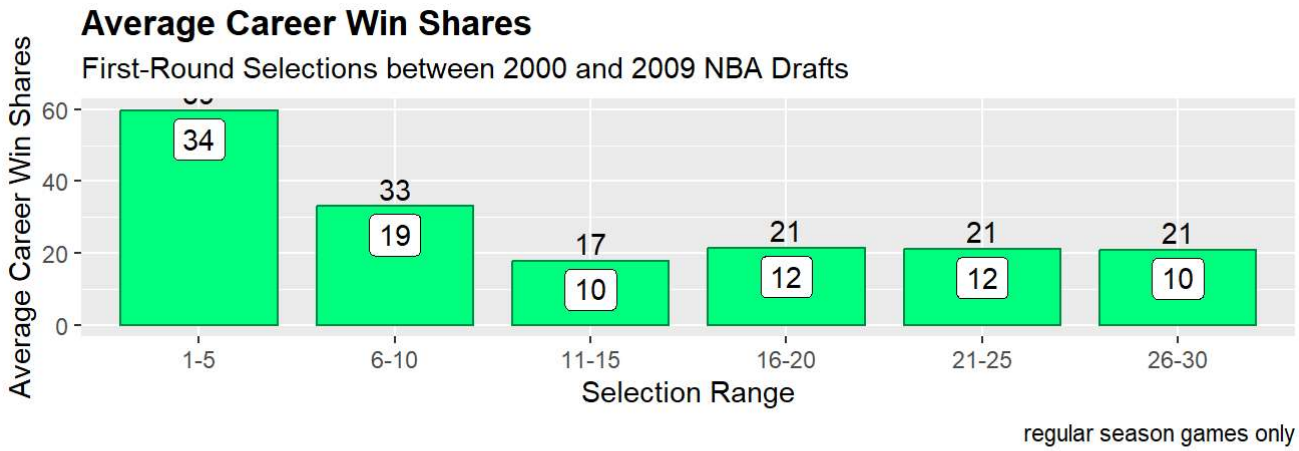
1.Players selected at or very near the top of the NBA draft averaged approximately 18% more minutes per game than players in the 6-10 selection range and almost twice as many minutes as players across the remaining selection ranges.

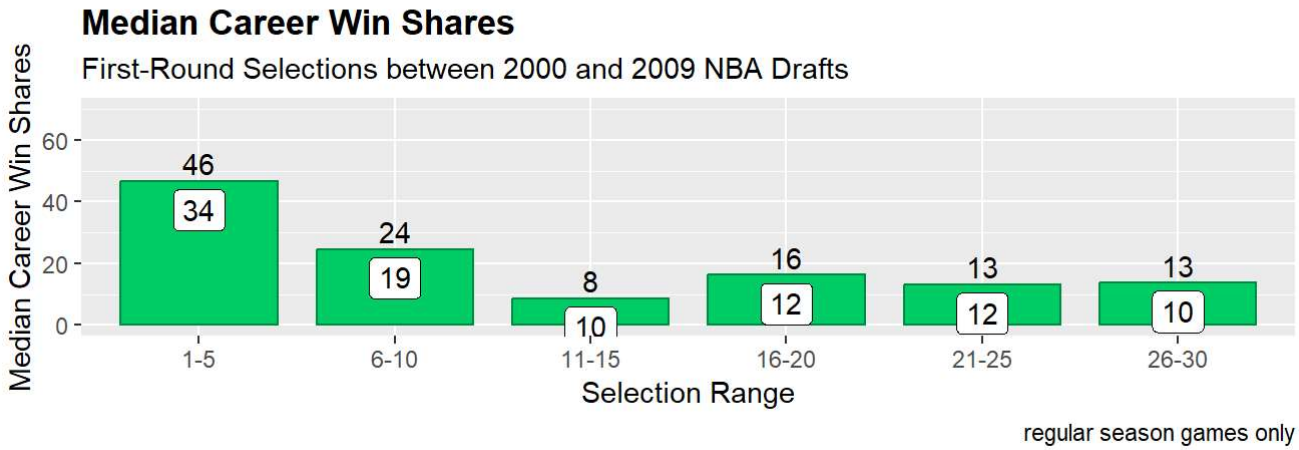
2.There is clear and obvious separation between selection ranges 1-5 and 6-10, and further separation between 6-10 and all other selection ranges.

3.There is almost no variance between selection ranges 11-15, 16-20, 21-25, and 26-30.

Win share comparesion

```
sumWS <- sum(draft2$WS)

tibble3 <- draft2 %>%
  dplyr::group_by(Pk2) %>%
  dplyr::summarize(mean = mean(WS),
            median = median(WS),
            pct = sum(WS)/sumWS)

ws1 <- ggplot(tibble3, aes(x = Pk2, y = mean)) +
  geom_bar(stat = "identity", width = .8, fill = "springgreen", color = "springgreen4") +
  labs(title = "Average Career Win Shares",
       subtitle = "First-Round Selections between 2000 and 2009 NBA Drafts",
       x = "Selection Range", y = "Average Career Win Shares",
       caption = "regular season games only") +
    scale_x_discrete(limits = c("1-5", "6-10", "11-15", "16-20", "21-25", "26-30"),
                     labels = c("1-5", "6-10", "11-15", "16-20", "21-25", "26-30")) +
  geom_text(aes(label = trunc(mean), vjust = -0.3)) +
  geom_label(aes(label = trunc(pct*100), vjust = 1.2)) +
  ylim(0, 60) +
  theme(plot.title = element_text(face = "bold"))

ws2 <- ggplot(tibble3, aes(x = Pk2, y = median)) +
  geom_bar(stat = "identity", width = .8, fill = "springgreen3", color = "springgreen4") +
  labs(title = "Median Career Win Shares",
       subtitle = "First-Round Selections between 2000 and 2009 NBA Drafts",
       x = "Selection Range", y = "Median Career Win Shares",
       caption = "regular season games only") +
    scale_x_discrete(limits = c("1-5", "6-10", "11-15", "16-20", "21-25", "26-30"),
                     labels = c("1-5", "6-10", "11-15", "16-20", "21-25", "26-30")) +
  geom_text(aes(label = trunc(median), vjust = -0.3)) +
  geom_label(aes(label = trunc(pct*100), vjust = 1.2)) +
  ylim(0, 70) +
  theme(plot.title = element_text(face = "bold"))

ws1 + ws2 + plot_layout(ncol = 1)
```





Key findings:

1.Players selected first through fifth accrued, on average, almost twice as many career win shares as those players then selected sixth through tenth and nearly three times as many career win shares as those players picked between the 11 spot and the end of the first round.

2.These same players–again, 17% of the draft2 population–account for 34% of all win shares; and players in the top-two selection ranges, roughly 34% of the draft2 population, account for at least 53% of all win shares.

3.The differences between the 1-5 selection range versus the 6-10 selection range versus the remaining selection ranges–or Pk2 levels–are more clear and more obvious than they were with respect to career games played or minutes played per game.

4.While there are significant differences at the top-end of the draft, there is little to no difference between selection range 11-15 all the way down to selection range 26-30.

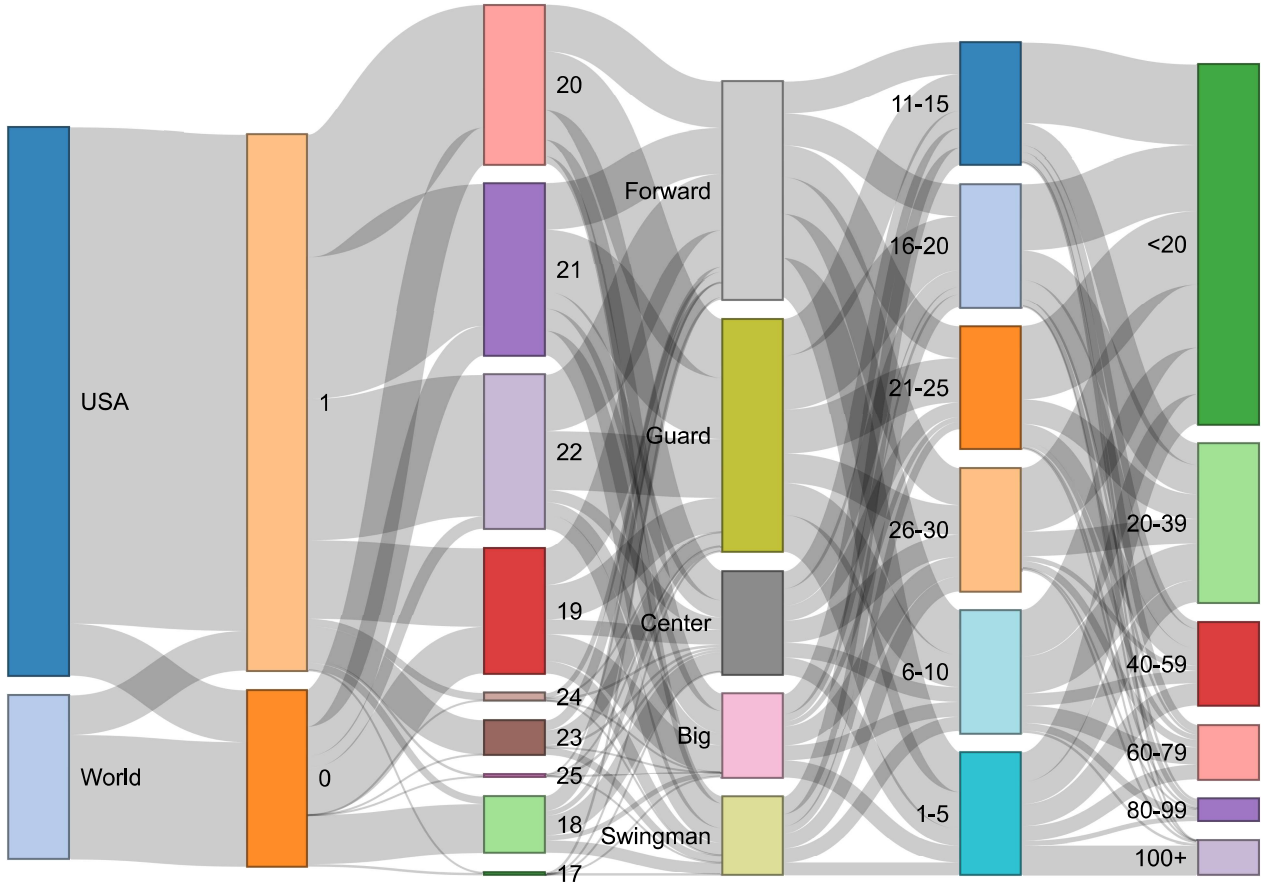
conclusions

Regardless of the metric, there is clear separation in performance between those players picked within the 1-5 selection range versus those picked in the 6-10 selection range; and there is further separation between the 6-10 selection range and those players then picked in the 11-15 selection range. From there, it doesn’t appear to matter. Teams are definitely best positioned to acquire top talent when picking higher in the draft; otherwise, it makes sense to trade down, since there is almost no difference in selection ranges 11-15 and 26-30. ## sankey graph

```
mutate(draft2, Age2 = trunc(Age)) -> draft2

mutate(draft2, WS2 = trunc(WS)) %>%
mutate(WS3 = case_when(WS2 <= 19 ~ "<20",
                           WS2 >= 20 & WS2 <= 39 ~ "20-39",
                           WS2 >= 40 & WS2 <= 59 ~ "40-59",
                           WS2 >= 60 & WS2 <= 79 ~ "60-79",
                           WS2 >= 80 & WS2 <= 99 ~ "80-99",
                           WS2 >= 100 ~ "100+")) -> draft2

nodes <- data.frame("name" = c("USA", "World",
                               "0", "1",
                               "17", "18", "19", "20", "21", "22", "23", "24", "25",
                               "Big",  "Center", "Forward", "Guard",  "Swingman",
                               "1-5",  "6-10", "11-15", "16-20", "21-25", "26-30",
                               "<20", "20-39", "40-59", "60-79", "80-99", "100+"))
links <- as.data.frame(matrix(c(
  0,2,21, 0,3,203,
  1,2,51, 1,3,16,
  2,4,1, 2,5,20, 2,6,19, 2,7,15, 2,8,12, 2,9,5, 2,10,0, 2,11,0, 2,12,0,
  3,4,0, 3,5,3, 3,6,32, 3,7,50, 3,8,58, 3,9,58, 3,10,14, 3,11,3, 3,12,1,
  4,13,0, 4,14,0, 4,15,1, 4,16,0, 4,17,0,
  5,13,2, 5,14,8, 5,15,6, 5,16,2, 5,17,5,
  6,13,11, 6,14,6, 6,15,15, 6,16,14, 6,17,5,
  7,13,7, 7,14,12, 7,15,19, 7,16,24, 7,17,3,
  8,13,9, 8,14,7, 8,15,19, 8,16,25, 8,17,10,
  9,13,5, 9,14,5, 9,15,23, 9,16,24, 9,17,6,
  10,13,0, 10,14,1, 10,15,4, 10,16,6, 10,17,3,
  11,13,0, 11,14,1, 11,15,2, 11,16,0, 11,17,0,
  12,13,0, 12,14,1, 12,15,0, 12,16,0, 12,17,0,
  13,18,7, 13,19,6, 13,20,8, 13,21,3, 13,22,2, 13,23,8,
  14,18,7, 14,19,6, 14,20,7, 14,21,7, 14,22,6, 14,23,9,
  15,18,16, 15,19,18, 15,20,13, 15,21,13, 15,22,13, 15,23,15,
  16,18,15, 16,19,13, 16,20,15, 16,21,22, 16,22,18, 16,23,12,
  17,18,5, 17,19,6, 17,20,7, 17,21,5, 17,22,3, 17,23,6,
  18,24,12, 18,25,9, 18,26,9, 18,27,6, 18,28,2, 18,29,12,
  19,24,19, 19,25,15, 19,26,5, 19,27,7, 19,28,3, 19,29,1,
  20,24,33, 20,25,9, 20,26,3, 20,27,3, 20,28,1, 20,29,0,
```

```
  21,24,27, 21,25,12, 21,26,8, 21,27,1, 21,28,2, 21,29,0,
  22,24,30, 22,25,10, 22,26,7, 22,27,2, 22,28,1, 22,29,0,
  23,24,26, 23,25,10, 23,26,2, 23,27,3, 23,28,0, 23,29,1),
  byrow = TRUE, ncol = 3))
names(links) = c("source", "target", "value")
```

```
sankeyNetwork(Links = links, Nodes = nodes,
                Source = "source", Target = "target",
                Value = "value", NodeID = "name",
                fontSize = 12, nodeWidth = 30)
```



Key findings:

Approximately four times as many first-round picks between the 2000 and 2009 NBA drafts were born in the United States versus some other country. Most first-round picks born in the US played in college before turning professional while a majority of first-round picks born outside the US did not play in college.

More first-round picks, regardless of where they were born and whether or not they first played in college, were aged 19-22 when they entered the NBA draft; very few players were either younger than 19 or older than 22 when they turned professional.

More first-round picks were forwards or guards than any other position.

A large majority of the players with 100 or more win shares in their respective careers were selected at or near the very top of the draft.

Most players with somewhere between 80 and 99 career win shares were selected between the 1 and 10 spots.

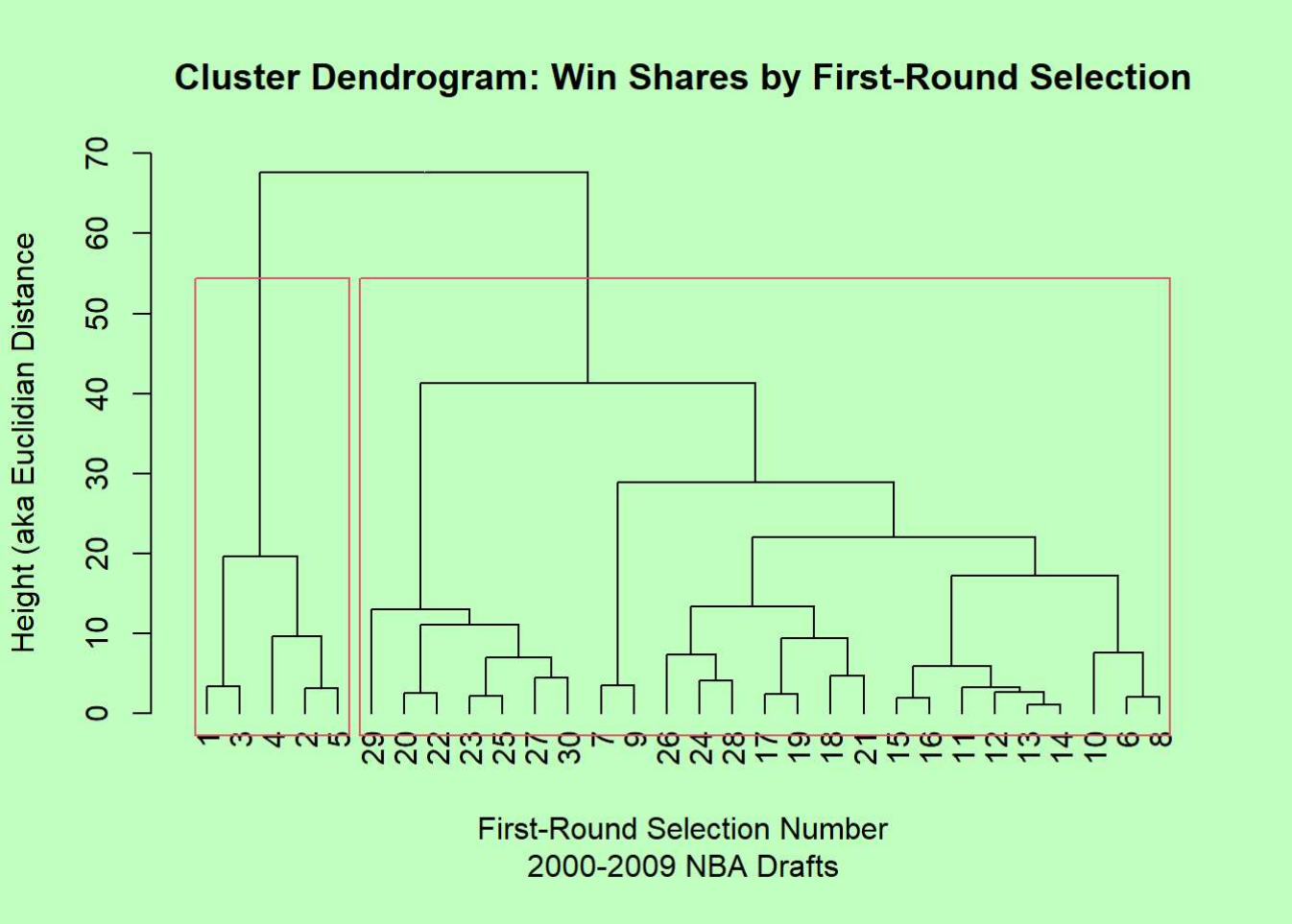
Hierarchical clustering

```
draft_clust <- select(draft2, Pk, WS)
draft_clust %>%
  group_by(Pk) %>%
  summarize(ws = mean(WS)) -> draft_clust_final
head(draft_clust_final)
```

Pk <int>	ws <dbl>
1	69.64
2	51.49
3	66.91
4	59.99
5	50.39
6	25.26

6 rows

```
d <- dist(draft_clust_final, method = "euclidean")
hc <- hclust(d, method = "complete")
bg = par(bg = "darkseagreen1")
plot(as.dendrogram(hc, cex = 0.6, hang = -1),
     main = "Cluster Dendrogram: Win Shares by First-Round Selection",
     xlab = "First-Round Selection Number\n2000-2009 NBA Drafts",
     ylab = "Height (aka Euclidian Distance")
rect.hclust(hc, k = 2)
```



Summary

Regardless of method, it’s clear that not all first-round picks are created equal. There is clear and obvious separation between players drafted at or near the very top of any NBA draft versus most every other available player.

Acquiring superstar talent is an absolute necessity for teams wanting to build a championship-caliber roster.

Therefore, teams must possess one of the first few picks to have a reasonable chance of selecting a potential superstar that can ultimately lead them to a championship.

Even with the lottery, about the only chance of selecting a superstar is by having a losing record during the prior season; better yet, teams should tank to ensure their record is worse than other teams.

We demonstrated the versatility of R by calling a combination of base and packaged functions to wrangle data and other combinations of base and packaged functions to compute and visualize results in traditional and not-so traditional ways.