# NBAProject

Kuanyu Lai

2023-04-29

```
#read data
draft <- read.csv("statisticsplaybook-main/statisticsplaybook-main/draft.csv",
                  header = TRUE, stringsAsFactors = FALSE)
```

```
dim(draft)
```

```
## [1] 293  26
```

```
glimpse(draft)
```

```
## Rows: 293
## Columns: 26
## $ 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, 20~
## $ Lg      <chr> "NBA", "NBA", "NBA", "NBA", "NBA", "NBA", "NBA", "NBA", "NBA",~
## $ Rd      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ 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 Ev~
## $ Age     <dbl> 20.106, 22.135, 19.308, 19.284, 18.252, 20.144, 21.108, 21.339~
## $ 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", "u~
## $ College <chr> "Oklahoma", "UConn", "Arizona State", "Memphis", NA, "Syracuse~
## $ From    <int> 2011, 2010, 2010, 2010, 2012, 2010, 2010, 2010, 2010, 2010, 20~
## $ To      <int> 2020, 2014, 2020, 2019, 2020, 2012, 2020, 2017, 2020, 2018, 20~
## $ G       <int> 622, 224, 826, 594, 555, 163, 699, 409, 813, 555, 153, 535, 42~
## $ MP      <dbl> 34.8, 10.5, 34.3, 30.7, 30.9, 22.9, 34.3, 18.8, 34.2, 30.3, 19~
## $ PTS     <dbl> 21.7, 2.2, 25.1, 15.7, 11.3, 9.2, 23.5, 7.9, 20.0, 14.1, 7.1, ~
## $ TRB     <dbl> 8.8, 2.7, 5.3, 4.6, 4.2, 1.9, 4.5, 5.8, 4.4, 3.0, 3.6, 3.2, 4.~
## $ AST     <dbl> 4.4, 0.1, 6.3, 4.8, 7.8, 3.9, 6.6, 0.8, 3.6, 5.7, 2.4, 1.9, 0.~
## $ STL     <dbl> 0.9, 0.3, 1.6, 1.2, 1.9, 0.7, 1.7, 0.4, 1.0, 1.2, 0.5, 0.7, 0.~
## $ BLK     <dbl> 0.5, 0.8, 0.5, 0.4, 0.1, 0.0, 0.2, 0.7, 0.3, 0.2, 0.1, 0.3, 0.~
## $ FG.     <dbl> 0.498, 0.567, 0.442, 0.440, 0.391, 0.400, 0.476, 0.497, 0.457,~
## $ X2P.    <dbl> 0.521, 0.567, 0.509, 0.468, 0.416, 0.420, 0.515, 0.499, 0.475,~
## $ X3P.    <dbl> 0.333, NA, 0.363, 0.323, 0.326, 0.338, 0.435, 0.136, 0.283, 0.~
## $ FT.     <dbl> 0.694, 0.578, 0.858, 0.757, 0.840, 0.809, 0.906, 0.699, 0.829,~
## $ 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, 0.115~
```

```
summary(draft)
```

```
##        Rk            Year           Lg                  Rd           Pk
##  Min.   :  1   Min.   :2000   Length:293         Min.   :1   Min.   : 1.00
##  1st Qu.: 74   1st Qu.:2002   Class :character   1st Qu.:1   1st Qu.: 8.00
##  Median :147   Median :2005   Mode  :character   Median :1   Median :15.00
##  Mean   :147   Mean   :2005                      Mean   :1   Mean   :15.16
##  3rd Qu.:220   3rd Qu.:2007                      3rd Qu.:1   3rd Qu.:22.00
##  Max.   :293   Max.   :2009                      Max.   :1   Max.   :30.00
##
##       Tm                Player               Age            Pos
##  Length:293         Length:293         Min.   :17.25   Length:293
##  Class :character   Class :character   1st Qu.:19.33   Class :character
##  Mode  :character   Mode  :character   Median :21.01   Mode  :character
##                                        Mean   :20.71
##                                        3rd Qu.:22.05
##                                        Max.   :25.02
##
##      Born              College               From            To
##  Length:293         Length:293         Min.   :2001   Min.   :2002
##  Class :character   Class :character   1st Qu.:2003   1st Qu.:2010
##  Mode  :character   Mode  :character   Median :2006   Median :2015
##                                        Mean   :2006   Mean   :2014
##                                        3rd Qu.:2008   3rd Qu.:2018
##                                        Max.   :2013   Max.   :2020
##                                        NA's   :2      NA's   :2
##        G                MP              PTS              TRB
##  Min.   :   6.0   Min.   : 4.30   Min.   : 0.700   Min.   : 0.500
##  1st Qu.: 248.0   1st Qu.:15.60   1st Qu.: 5.350   1st Qu.: 2.200
##  Median : 549.0   Median :21.60   Median : 8.000   Median : 3.300
##  Mean   : 526.4   Mean   :21.53   Mean   : 8.869   Mean   : 3.779
##  3rd Qu.: 789.5   3rd Qu.:27.70   3rd Qu.:11.250   3rd Qu.: 4.850
##  Max.   :1326.0   Max.   :38.40   Max.   :27.100   Max.   :12.300
##  NA's   :2        NA's   :2       NA's   :2        NA's   :2
##       AST              STL              BLK              FG.
##  Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.700   1st Qu.:0.400   1st Qu.:0.2000   1st Qu.:0.4125
##  Median :1.300   Median :0.600   Median :0.3000   Median :0.4400
##  Mean   :1.849   Mean   :0.644   Mean   :0.4732   Mean   :0.4448
##  3rd Qu.:2.300   3rd Qu.:0.800   3rd Qu.:0.6000   3rd Qu.:0.4800
##  Max.   :9.500   Max.   :2.200   Max.   :2.1000   Max.   :0.6070
##  NA's   :2       NA's   :2       NA's   :2        NA's   :2
##       X2P.             X3P.              FT.              WS
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.390   Min.   : -1.60
##  1st Qu.:0.4450   1st Qu.:0.2350   1st Qu.:0.672   1st Qu.:  4.05
##  Median :0.4700   Median :0.3275   Median :0.744   Median : 19.60
##  Mean   :0.4681   Mean   :0.2768   Mean   :0.729   Mean   : 29.35
##  3rd Qu.:0.4995   3rd Qu.:0.3580   3rd Qu.:0.800   3rd Qu.: 43.85
##  Max.   :0.6100   Max.   :1.0000   Max.   :1.000   Max.   :236.10
##  NA's   :2        NA's   :13       NA's   :2       NA's   :2
##      WS.48
##  Min.   :-0.32600
##  1st Qu.: 0.05000
```

```
##  Median : 0.07900
##  Mean   : 0.07592
##  3rd Qu.: 0.10600
##  Max.   : 0.24400
##  NA's   :2
```

```
#removing variables
draft <- select(draft, -c(3,4,16:24))
draft <- draft[-c(90, 131),]
```

```
head(draft, 10)
```

```
##    Rk Year Pk  Tm           Player    Age Pos Born       College From   To   G
## 1   1 2009  1 LAC    Blake Griffin 20.106   F   us      Oklahoma 2011 2020 622
## 2   2 2009  2 MEM  Hasheem Thabeet 22.135   C   tz         UConn 2010 2014 224
## 3   3 2009  3 OKC     James Harden 19.308   G   us Arizona State 2010 2020 826
## 4   4 2009  4 SAC     Tyreke Evans 19.284 G-F   us       Memphis 2010 2019 594
## 5   5 2009  5 MIN      Ricky Rubio 18.252   G   es          <NA> 2012 2020 555
## 6   6 2009  6 MIN      Jonny Flynn 20.144   G   us      Syracuse 2010 2012 163
## 7   7 2009  7 GSW    Stephen Curry 21.108   G   us      Davidson 2010 2020 699
## 8   8 2009  8 NYK      Jordan Hill 21.339 C-F   us       Arizona 2010 2017 409
## 9   9 2009  9 TOR    DeMar DeRozan 19.327 G-F   us           USC 2010 2020 813
## 10 10 2009 10 MIL Brandon Jennings 19.280   G   us          <NA> 2010 2018 555
##       MP    WS  WS.48
## 1   34.8  75.2  0.167
## 2   10.5   4.8  0.099
## 3   34.3 133.3  0.226
## 4   30.7  28.4  0.075
## 5   30.9  36.4  0.102
## 6   22.9  -1.1 -0.015
## 7   34.3 103.2  0.207
## 8   18.8  16.4  0.102
## 9   34.2  66.5  0.115
## 10 30.3  29.9  0.085
```

```
#convert data type
draft$Year <- as.factor(draft$Year)
draft$Tm <- as.factor(draft$Tm)
draft$Pos <- as.factor(draft$Pos)
draft$Born <- as.factor(draft$Born)
draft$From <- as.factor(draft$From)
draft$To <- as.factor(draft$To)
```

```
#changing Born to either USA or WORLD
draft <- mutate(draft, born2 = ifelse(Born == "us", "USA", "World"))
draft$born2 <- as.factor(draft$born2)
```

```
#dealing with 0
draft[is.na(draft)] <- 0
mutate(draft, College2 = ifelse(College == 0, 0, 1)) -> draft
draft$College2 <- factor(draft$College2)
```

```
#rename positions
levels(draft$Pos)
```

```
## [1] "C"   "C-F" "F"   "F-C" "F-G" "G"   "G-F"
```

```
mutate(draft, Pos2 = ifelse(Pos == "G", "Guard",
                     ifelse(Pos == "F", "Forward",
                     ifelse(Pos == "C", "Center",
                     ifelse(Pos == "F-G", "Swingman",
                     ifelse(Pos == "G-F", "Swingman",
                     ifelse(Pos == "F-C", "Big",
                     ifelse(Pos == "C-F", "Big", "NA"))))))) -> draft
```

```
summary(draft)
```

```
##        Rk             Year           Pk              Tm         Player
##  Min.   :  1.0   2006   : 30   Min.   : 1.00   BOS    : 13   Length:291
##  1st Qu.: 73.5   2008   : 30   1st Qu.: 8.00   CHI    : 13   Class :character
##  Median :148.0   2009   : 30   Median :15.00   POR    : 13   Mode  :character
##  Mean   :147.3   2000   : 29   Mean   :15.12   MEM    : 12
##  3rd Qu.:220.5   2003   : 29   3rd Qu.:22.00   NJN    : 12
##  Max.   :293.0   2004   : 29   Max.   :30.00   PHO    : 12
##                  (Other):114                   (Other):216
##       Age          Pos          Born       College               From
##  Min.   :17.25   C  :42   us     :224   Length:291         2005   : 31
##  1st Qu.:19.33   C-F:10   es     :  6   Class :character   2009   : 31
##  Median :21.01   F  :88   fr     :  6   Mode  :character   2002   : 30
##  Mean   :20.71   F-C:24   br     :  4                      2004   : 29
##  3rd Qu.:22.05   F-G:10   si     :  4                      2006   : 29
##  Max.   :25.02   G  :95   de     :  3                      2007   : 28
##                  G-F:22   (Other): 44                     (Other):113
##        To              G                MP              WS
##  2020   : 46   Min.   :   6.0   Min.   : 4.30   Min.   : -1.60
##  2019   : 24   1st Qu.: 248.0   1st Qu.:15.60   1st Qu.:  4.05
##  2013   : 23   Median : 549.0   Median :21.60   Median : 19.60
##  2017   : 23   Mean   : 526.4   Mean   :21.53   Mean   : 29.35
##  2015   : 22   3rd Qu.: 789.5   3rd Qu.:27.70   3rd Qu.: 43.85
##  2018   : 18   Max.   :1326.0   Max.   :38.40   Max.   :236.10
##  (Other):135
##      WS.48            born2     College2     Pos2
##  Min.   :-0.32600   USA  :224   0: 72    Length:291
##  1st Qu.: 0.05000   World: 67   1:219    Class :character
##  Median : 0.07900                        Mode  :character
##  Mean   : 0.07592
##  3rd Qu.: 0.10600
##  Max.   : 0.24400
##
```

Interesting Findings:

1.There is a tremendous amount of variance in career win shares. At least one first-round pick between the 2000 and 2009 NBA drafts actually accrued a negative number of win shares. And one player accrued more than 236 win shares.

2.In fact, there is also significant variances in the other career statistics, namely regular season games played and average minutes played per regular season game.

3.Going back to win shares, the mean, which is especially sensitive to outliers (or data points far removed from the population center), is significantly greater than the median, suggesting that the mean is skewed by a small number of superstars in the data set.

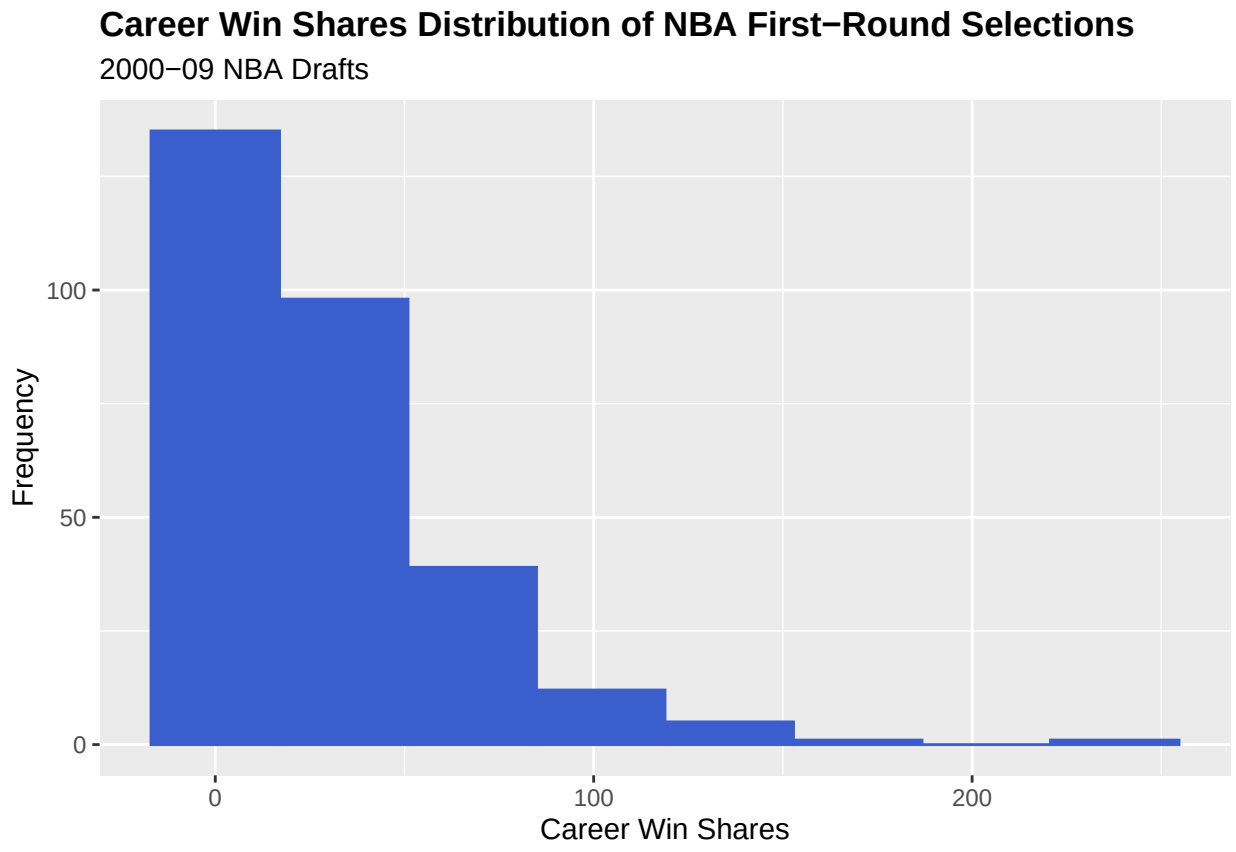
4.First-round NBA draft picks between 2000 and 2009 were anywhere between 17.25 years in age and 25.02 years at the time they were selected.

5.More than three-quarters of the players in draft, 224 of 291 to be exact, were born in the United States.

6.Nearly the same number of players, 219 to be specific, attended college or university.

### visualizing some data

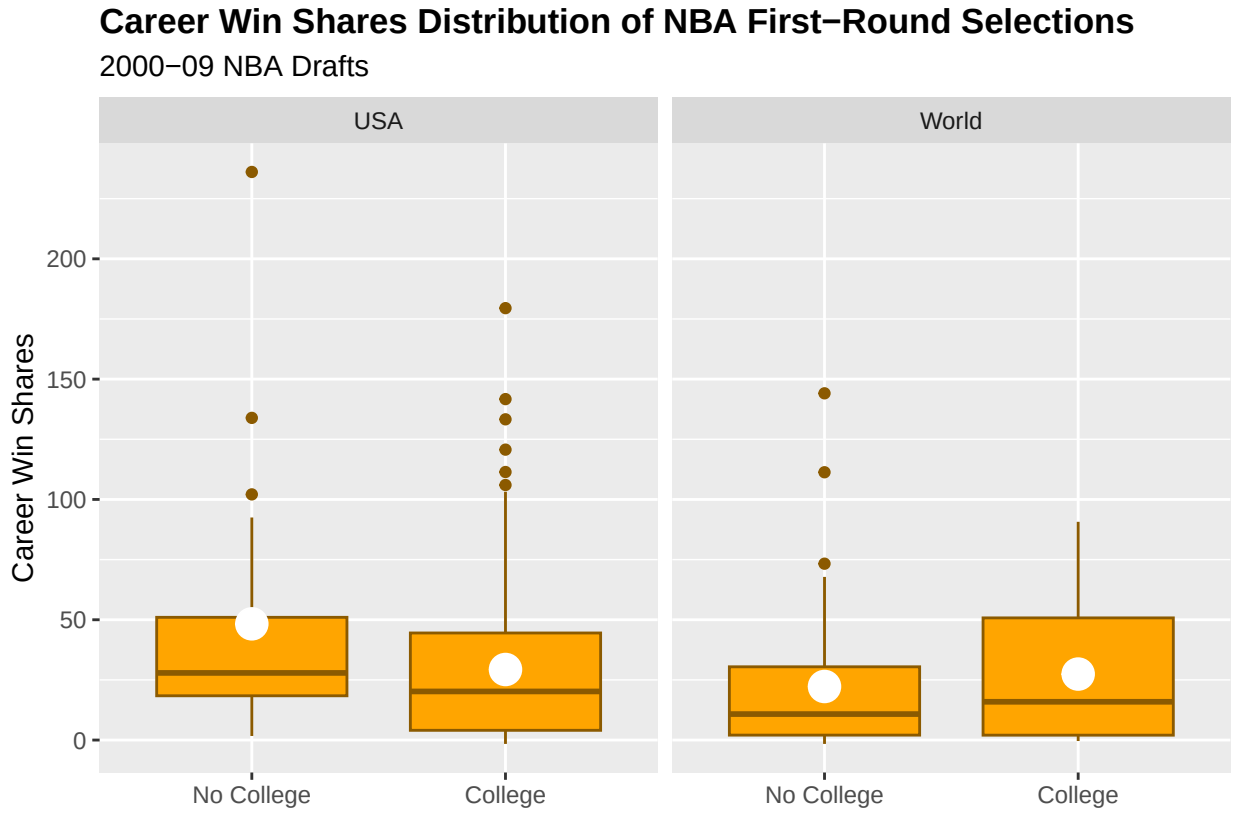
```
p1 <- ggplot(draft, aes(x = WS)) +
  geom_histogram(fill = "royalblue3", color = "royalblue3", bins = 8) +
  labs(title = "Career Win Shares Distribution of NBA First-Round Selections",
       subtitle = "2000-09 NBA Drafts",
       x = "Career Win Shares", y = "Frequency") +
  theme(plot.title = element_text(face = "bold"))
print(p1)
```




The variable win shares has a right-skewed, or positive-skewed, distribution–right-skewed because the distribution has a long right tail, positive-skewed because the long tail is in the positive direction of the

x-axis. In layman terms, it simply means that many NBA first-round picks between the 2000 and 2009 drafts accrued very few career win shares while just a few players accrued lots of win shares.

```
p2 <- ggplot(draft, aes(x = College2, y = WS)) +
  geom_boxplot(color = "orange4", fill = "orange1") +
  labs(title = "Career Win Shares Distribution of NBA First-Round Selections",
       x = "", y = "Career Win Shares", subtitle = "2000-09 NBA Drafts") +
  stat_summary(fun = mean, geom = "point", shape = 20, size = 8, color = "white", fill = "white") +
  theme(plot.title = element_text(face = "bold")) +
  facet_wrap(~born2) +
  scale_x_discrete(breaks = c(0, 1),
                        labels = c("No College", "College"))
print(p2)
```




Key findings:

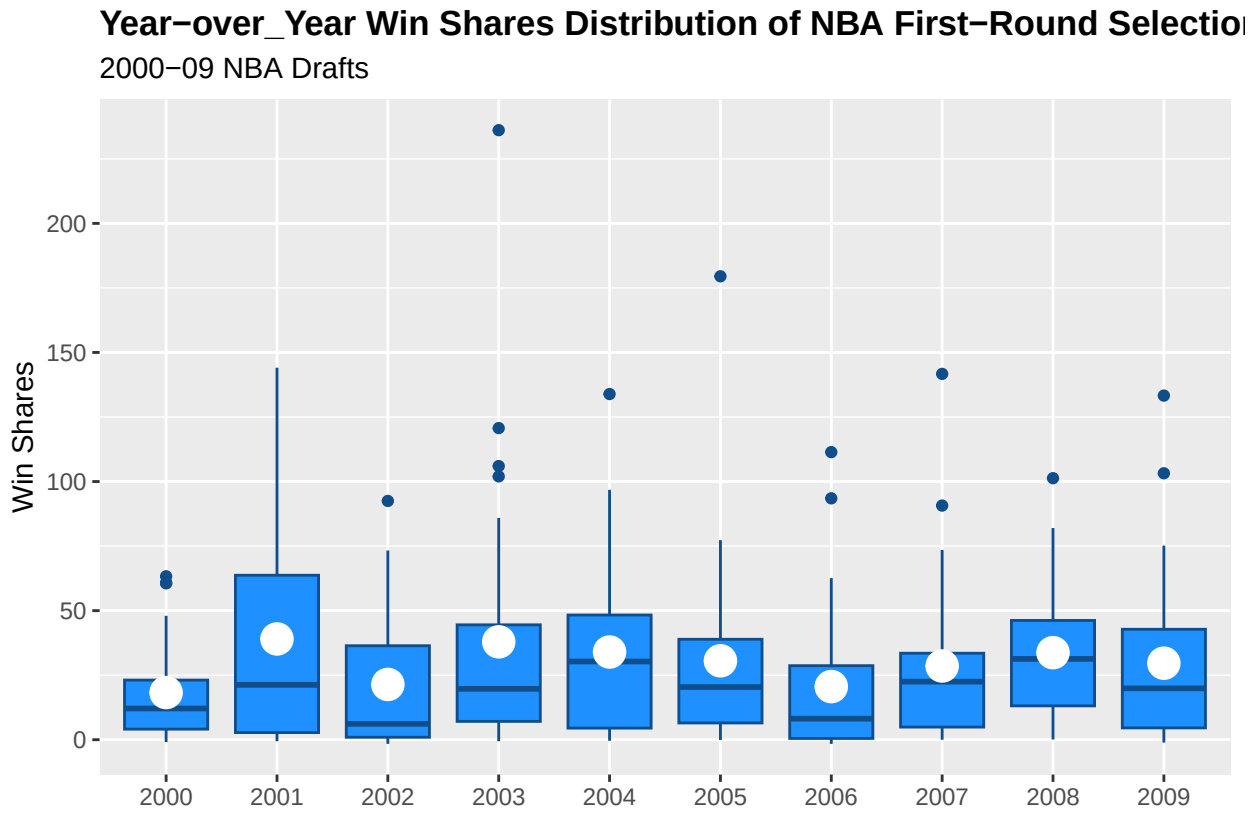
1.Players born in the United States generally accrued more career win shares than did players born outside the US.

2.Players born in the US who bypassed college generally accrued more career win shares than did US-born players who did attend a college or university.

3.Alternatively, players born in any country other than the United States and who did not attend college or university generally accrued fewer win shares over their respective careers than did other players born outside the US who did attend college.

4.The means are consistently higher than the medians, suggesting of course that average win shares, regardless of how the data are sliced and diced, are affected by 5.superstar production. More boxplots display the distribution of career win shares by year, or draft class

```
p3 <- ggplot(draft, aes(x = Year, y = WS)) +
  geom_boxplot(color = "dodgerblue4", fill = "dodgerblue" ) +
  labs(title = "Year-over_Year Win Shares Distribution of NBA First-Round Selections",
       x = "", y = "Win Shares", subtitle = "2000-09 NBA Drafts") +
  stat_summary(fun = mean, geom = "point", shape = 20, size = 8, color = "white", fill = "white") +
  theme(plot.title = element_text(face = "bold"))
print(p3)
```




```
draft %>%
  filter(WS >= 75) %>%
  group_by(Year) %>%
  summarize(avg = mean(Pk)) -> fourth_tibble
print(fourth_tibble)
```

```
## # A tibble: 9 x 2
##   Year    avg
##   <fct> <dbl>
## 1 2001   11.6
## 2 2002    9
## 3 2003    6.2
## 4 2004    5
## 5 2005   12.3
## 6 2006   13
## 7 2007    2.5
## 8 2008    4.5
```

```
## 9 2009   3.67
```

### Visualizing correlations

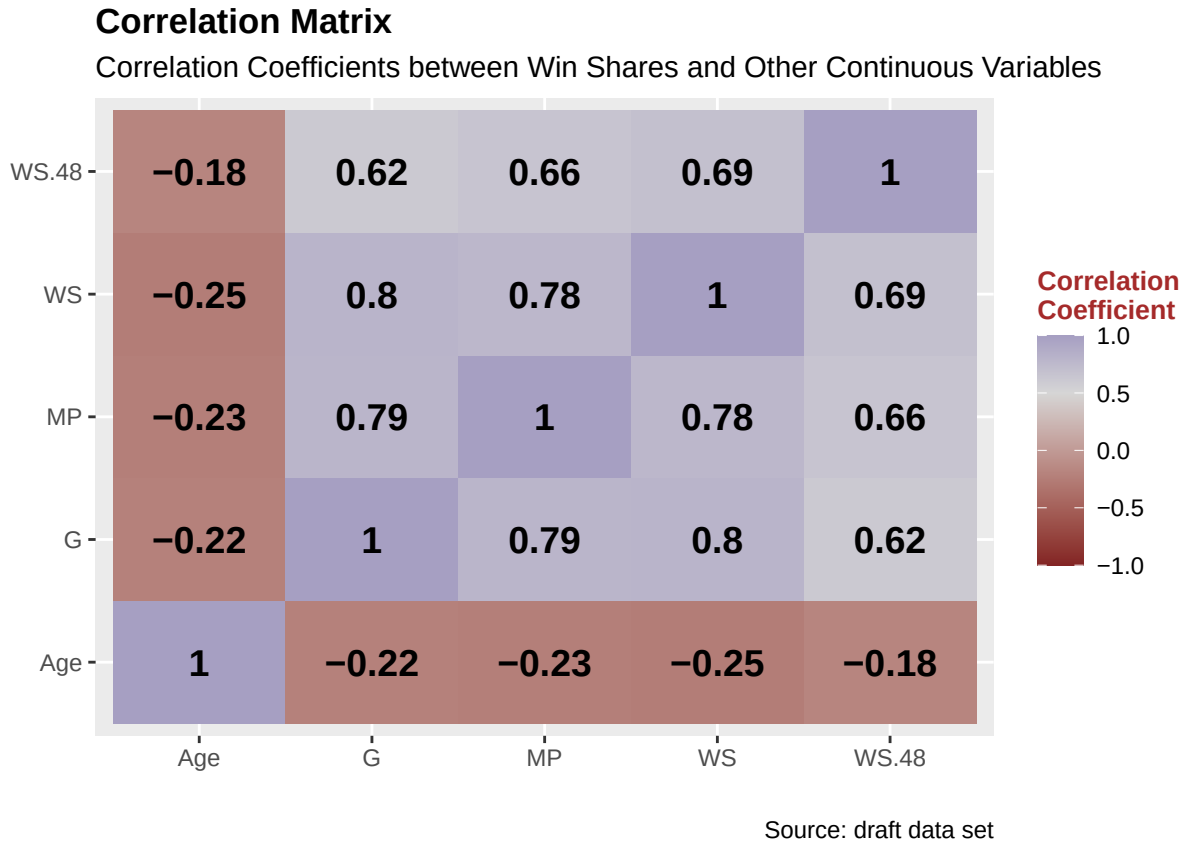
```
draft_cor <- select(draft, c("Age", "G","MP","WS","WS.48"))
draft_cor <- cor(draft_cor)
draft_cor
```

```
##              Age          G         MP         WS      WS.48
## Age    1.0000000 -0.2189601 -0.2327846 -0.2509647 -0.1801535
## G     -0.2189601  1.0000000  0.7921621  0.8004797  0.6165429
## MP    -0.2327846  0.7921621  1.0000000  0.7758876  0.6597869
## WS    -0.2509647  0.8004797  0.7758876  1.0000000  0.6942061
## WS.48 -0.1801535  0.6165429  0.6597869  0.6942061  1.0000000
```

```
draft_cor <- melt(draft_cor, na.rm = T)
head(draft_cor)
```

```
##    Var1 Var2      value
## 1   Age  Age  1.0000000
## 2     G  Age -0.2189601
## 3    MP  Age -0.2327846
## 4    WS  Age -0.2509647
## 5 WS.48  Age -0.1801535
## 6   Age    G -0.2189601
```

```
p4 <- ggplot(data =draft_cor , aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(midpoint = 0.5, mid = "grey84", limits = c(-1, 1)) +
  labs(title = "Correlation Matrix",
    subtitle = "Correlation Coefficients between Win Shares and Other Continuous Variables",
    x = "", y = "", fill = "Correlation\nCoefficient", caption = "Source: draft data set") +
  theme(plot.title = element_text(face = "bold"),
  legend.title = element_text(face = "bold", color = "brown", size = 10)) +
  geom_text(aes(x = Var1, y = Var2, label = round(value, 2)), color = "black",
  fontface = "bold", size = 5)
print(p4)
```




key findings:

There are positive, and strong, correlations between win shares and regular season games played, minutes played per regular season game, and win shares for every 48 minutes of playing time. Mind you, correlation coefficients don't tell us which variable might be influencing another variable, if in fact there is any causation at all.

However, there is a negative correlation between the variables Win Shares and Age; which is to say that players entering the NBA draft between 2000 and 2009 at younger ages then accrued, generally, more career win shares than players who turned professional at "older" ages. The correlation between these variables is not strong, however. No doubt this is partially true because younger players likely have more years to play professional ball, and therefore more opportunity to accrue more win shares; but it's also true–or at least likely so–that better players turn professional at younger ages than lesser players.

## Visualize mean and median

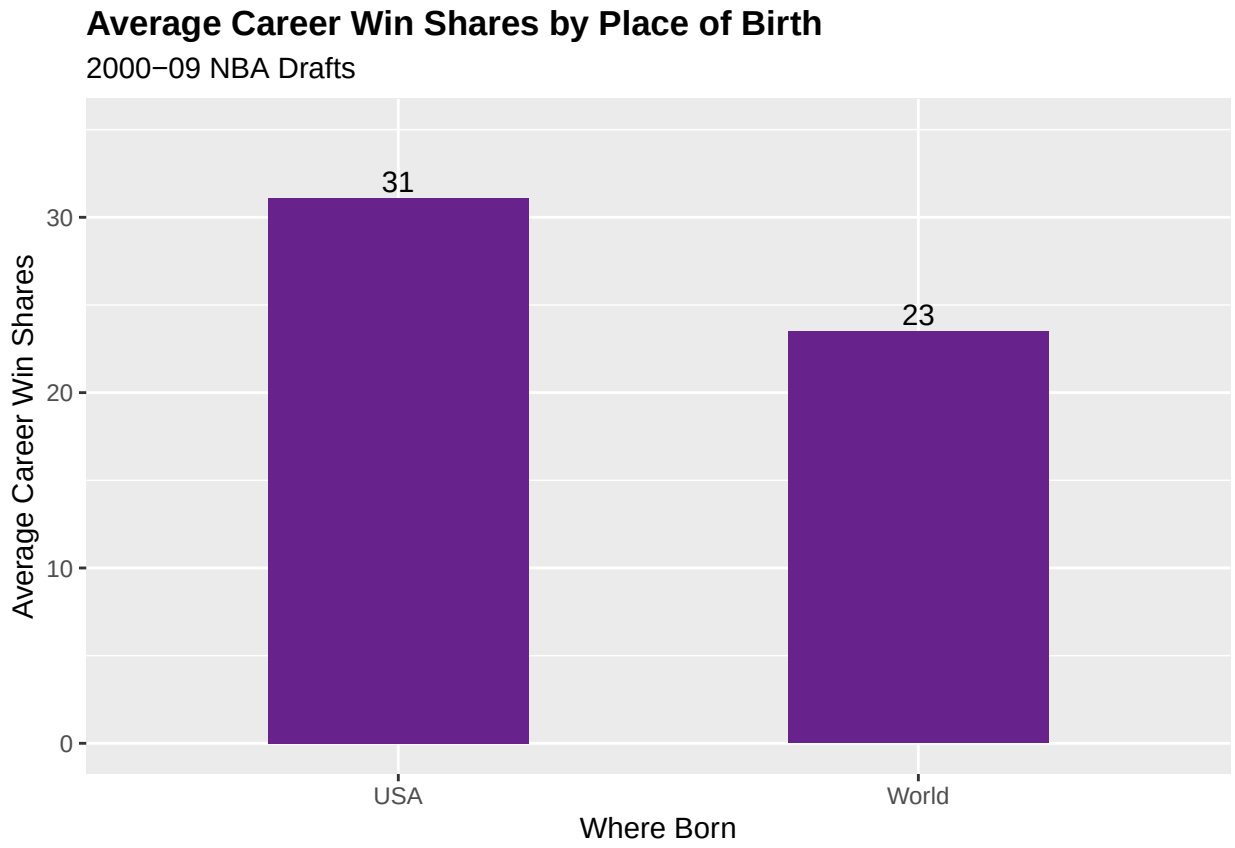
```
mean_median <- draft %>%
  dplyr::group_by(born2) %>%
  dplyr::summarize(meanWS = mean(WS),
            medianWS = median(WS))
```

```
  p5 <- ggplot(mean_median, aes(x = born2, y = meanWS)) +
  geom_bar(stat = "identity", width = .5, fill  = "darkorchid4") +
  labs(title = "Average Career Win Shares by Place of Birth",
       subtitle = "2000-09 NBA Drafts",
       x = "Where Born", y = "Average Career Win Shares") +
```

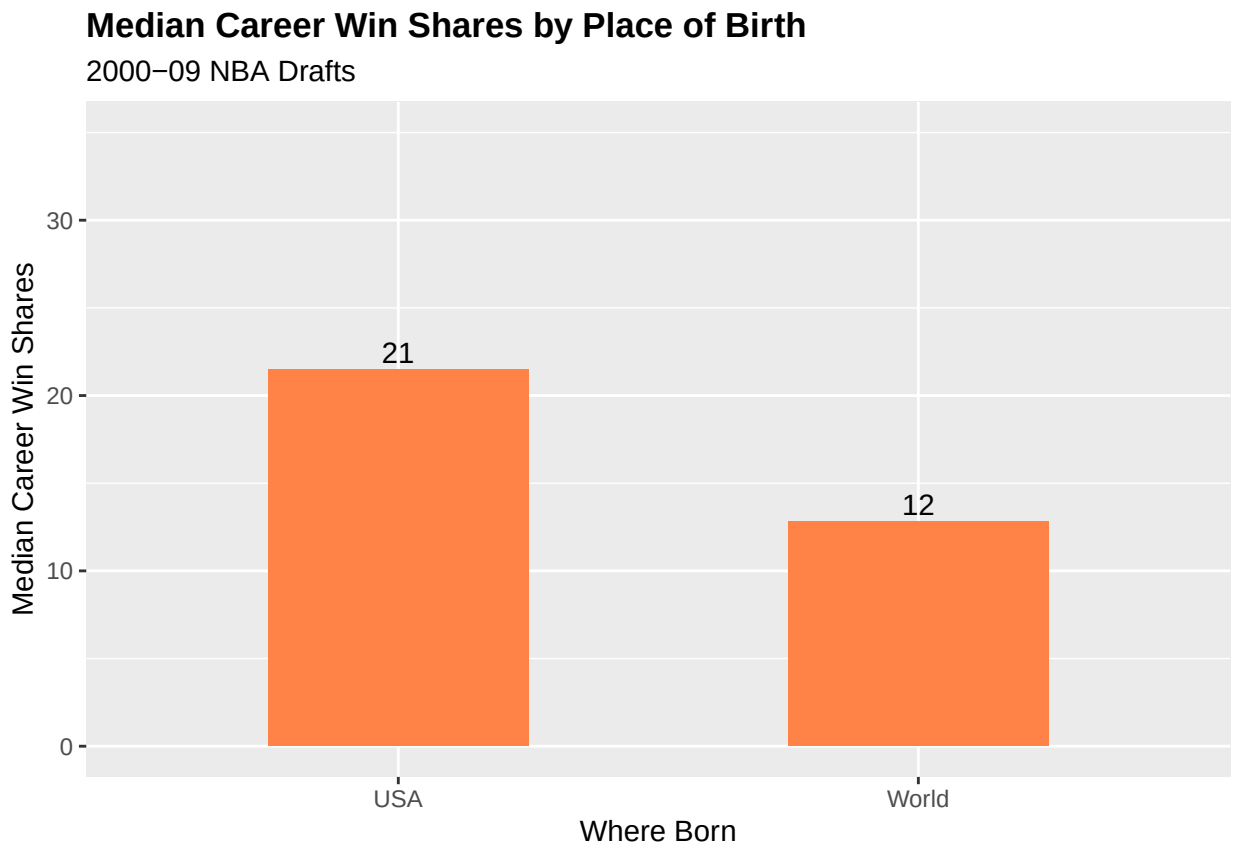
```
  geom_text(aes(label = trunc(meanWS), vjust = -0.3)) +
  ylim(0, 35) +
  theme(plot.title = element_text(face = "bold"))

p6 <- ggplot(mean_median, aes(x = born2, y = medianWS)) +
  geom_bar(stat = "identity", width = .5, fill  = "sienna1") +
  labs(title = "Median Career Win Shares by Place of Birth",
       subtitle = "2000-09 NBA Drafts",
       x = "Where Born", y = "Median Career Win Shares") +
  geom_text(aes(label = trunc(medianWS), vjust = -0.3)) +
  ylim(0, 35) +
  theme(plot.title = element_text(face = "bold"))

par(mfrow = c(1,2))
p5
```




```
p6
```




key Findings:

1.Players born in the US, on average, accrue more career win shares than players born elsewhere.

2.The means are significantly greater than the medians. Means are sensitive to outliers whereas medians are not; that the means are greater than the medians suggests, of course, that they are influenced by superstar production where win shares per superstar is greater than 100.

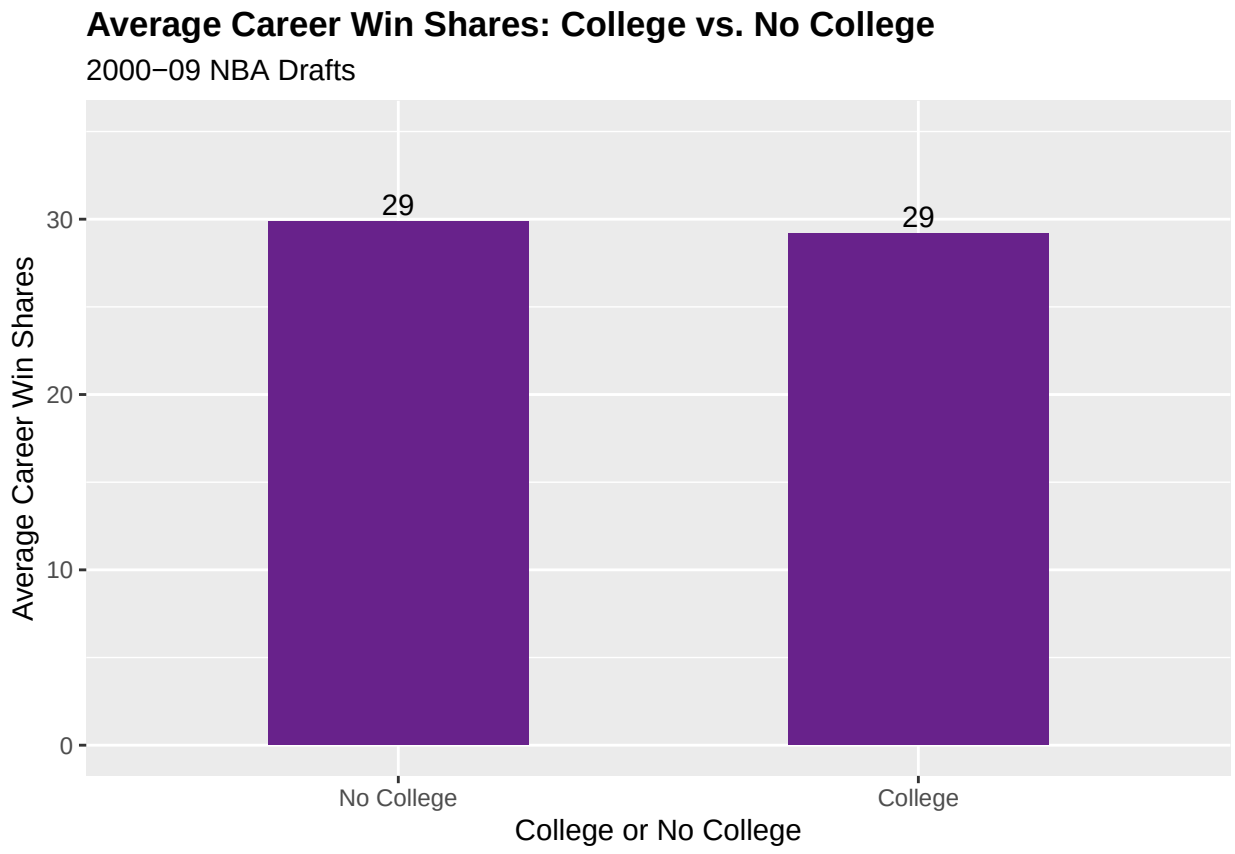
```
draft %>%
  dplyr::group_by(College2) %>%
  dplyr::summarize(meanWS = mean(WS),
            medianWS = median(WS)) -> ninth_tibble
print(ninth_tibble)
```

```
## # A tibble: 2 x 3
##   College2 meanWS medianWS
##   <fct>     <dbl>    <dbl>
## 1 0          29.9     19.2
## 2 1          29.2     20.1
```

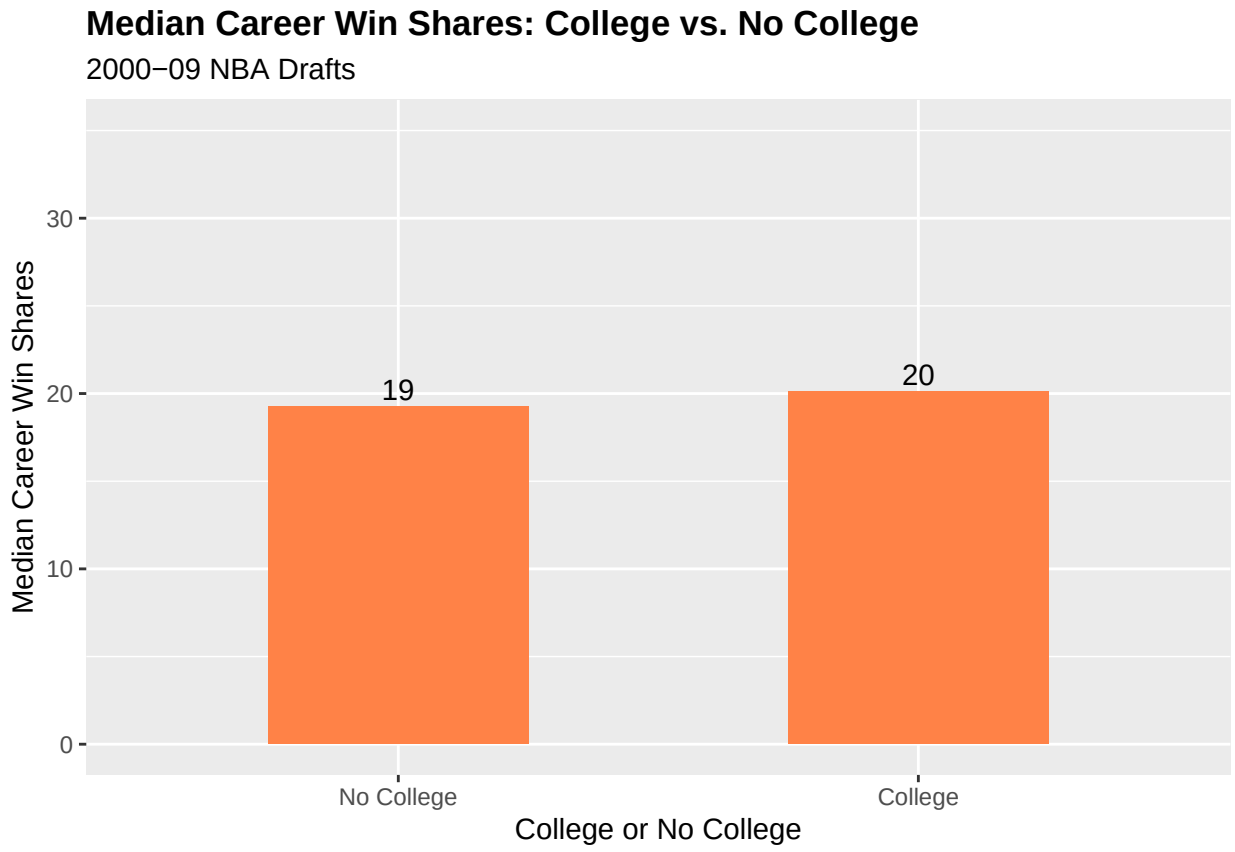
```
p7 <- ggplot(ninth_tibble, aes(x = College2, y = meanWS)) +
  geom_bar(stat = "identity", width = .5, fill  = "darkorchid4") +
  labs(title = "Average Career Win Shares: College vs. No College",
       subtitle = "2000-09 NBA Drafts",
       x = "College or No College", y = "Average Career Win Shares") +
  scale_x_discrete(breaks = c(0, 1),
                        labels = c("No College", "College")) +
```

```
  geom_text(aes(label = trunc(meanWS), vjust = -0.3)) +
  ylim(0, 35) +
  theme(plot.title = element_text(face = "bold"))

p8 <- ggplot(ninth_tibble, aes(x = College2, y = medianWS)) +
  geom_bar(stat = "identity", width = .5, fill  = "sienna1") +
  labs(title = "Median Career Win Shares: College vs. No College",
       subtitle = "2000-09 NBA Drafts",
       x = "College or No College", y = "Median Career Win Shares") +
  scale_x_discrete(breaks = c(0, 1),
                        labels = c("No College", "College")) +
  geom_text(aes(label = trunc(medianWS), vjust = -0.3)) +
  ylim(0, 35) +
  theme(plot.title = element_text(face = "bold"))
par(mfrow = c(1,2))
p7
```




```
p8
```

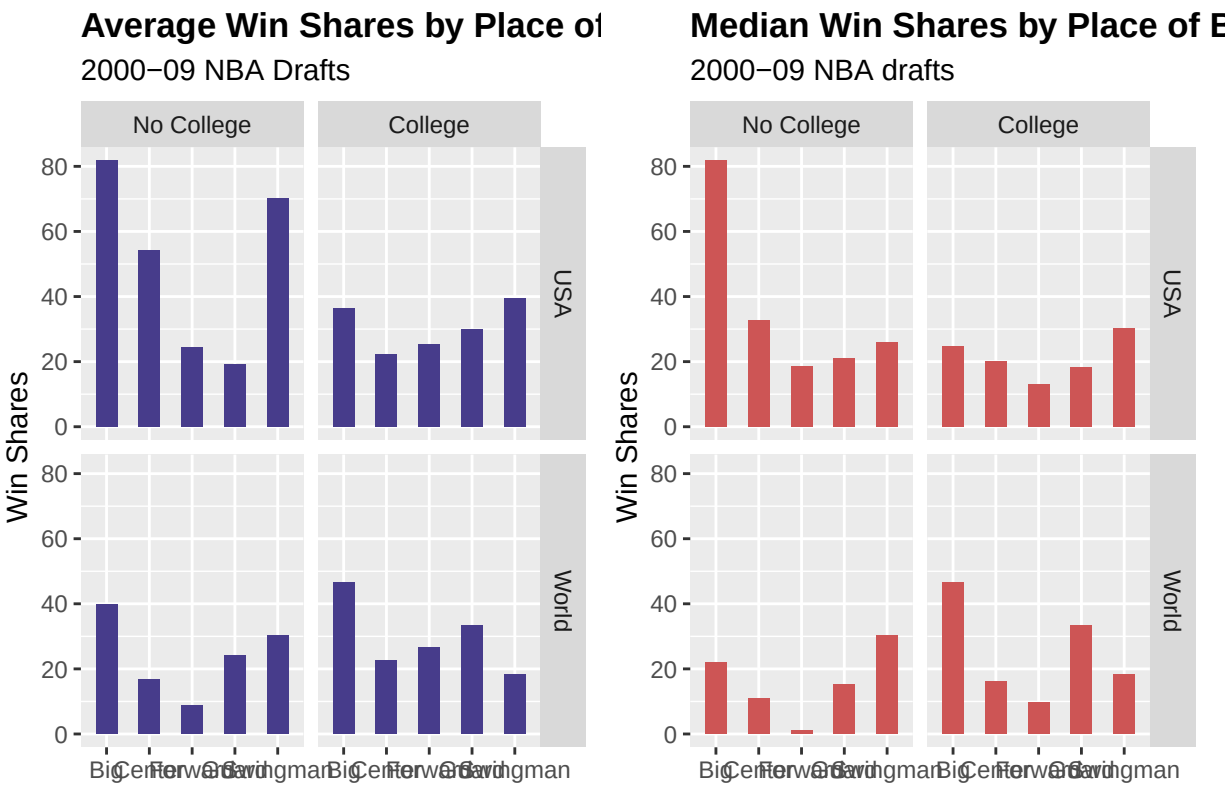



```
tenth_tibble <- draft %>%
                  dplyr::group_by(Pos2, born2, College2) %>%
                  dplyr::summarize(mean = mean(WS),
                                   median = median(WS))
```

```
## `summarise()` has grouped output by 'Pos2', 'born2'. You can override using the
## `.groups` argument.
```

```
new_labels <- c("0" = "No College", "1" = "College")
p9 <- ggplot(tenth_tibble, aes(x = Pos2, y = mean)) +
  geom_bar(stat = "identity", width = .5, fill  = "slateblue4") +
  labs(title = "Average Win Shares by Place of Birth", x = "", y = "Win Shares",
       subtitle = "2000-09 NBA Drafts") +
  theme(plot.title = element_text(face = "bold")) +
  facet_grid(born2 ~ College2, labeller = labeller(College2 = new_labels))

new_labels <- c("0" = "No College", "1" = "College")
p10 <- ggplot(tenth_tibble, aes(x = Pos2, y = median)) +
  geom_bar(stat = "identity", width = .5, fill  = "indianred3") +
  labs(title = "Median Win Shares by Place of Birth", x = "", y = "Win Shares",
       subtitle = "2000-09 NBA drafts") +
  theme(plot.title = element_text(face = "bold")) +
  facet_grid(born2 ~ College2, labeller = labeller(College2 = new_labels))
p9 + p10 + plot_layout(ncol = 2)
```




The most obvious, and fascinating, results found in our pair of facet plots, are in the upper-left panels. Bigs, Centers, and Swingmen born in the US who did not attend a college or university before entering the NBA draft accrued significantly more win shares, on average, than other players at other positions regardless of where they were born and regardless of whether or not they first attended college.

```
draft -> draft2
```