NCAA Basketball - March Madness

Stanton Kent

2024-10-02

The Perfect March Madness Bracket

How difficult is it? If you took a pure 50/50 chance across all 67 games, you would end up with a 1-in-9.2 quintillion. Not every game is a coinflip though. What if you knew a lot about basketball? The following uses the NCAA tournament data from 1985 - 2017 how easily different aspects of the simplest data can be used to estimate wins.

Setting up my environment

Notes: Setting up my environment by loading the 'tidyverse' package, along with a .csv file generated from the ncaa_basketball public dataset from BigQuery. The dataset is copyrighted by NCAA ands SportsTradar to be used for research purposes.

```
install.packages("tidyverse")
ncaa_pts <- read.csv("ncaa_pts.csv")
library(tidyverse)
```

Setting up the datasets

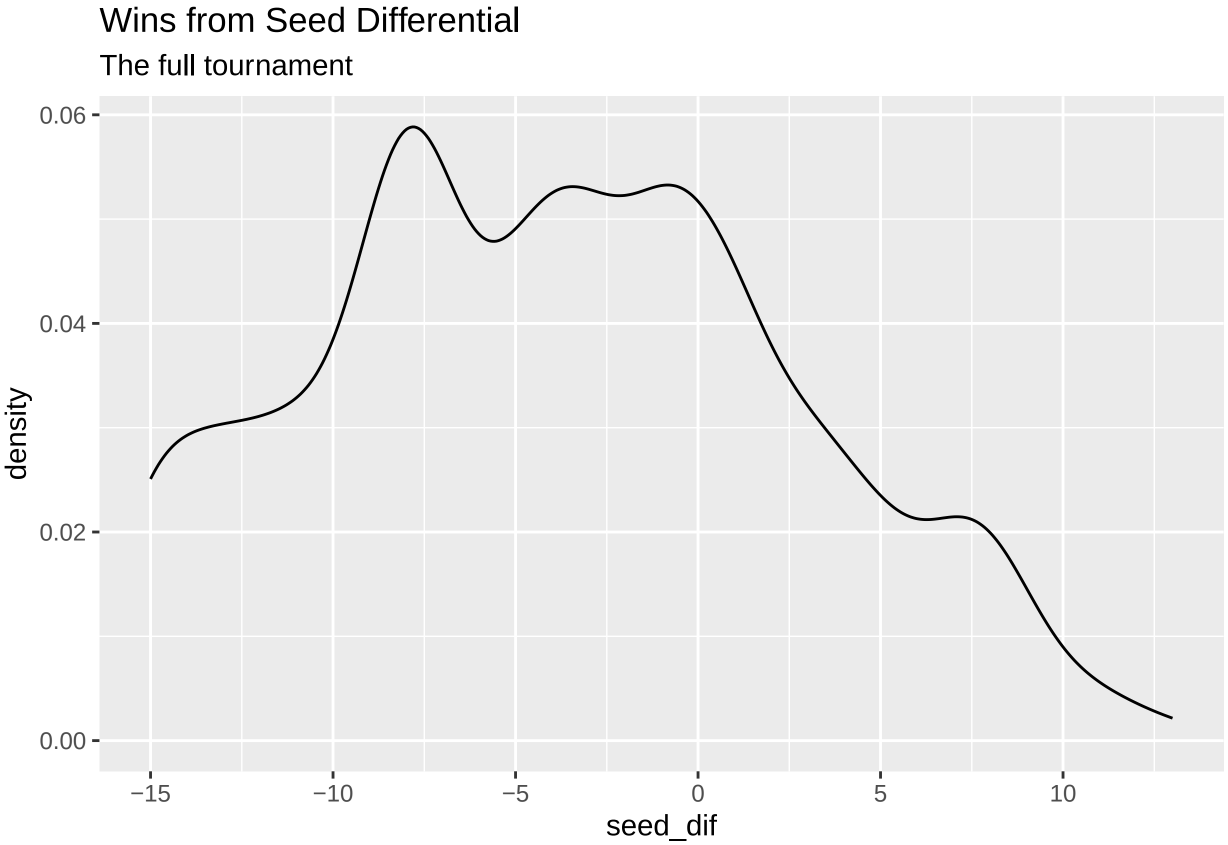
```
ncaa_pts2 <- mutate(ncaa_pts, pts_dif = win_pts - lose_pts)
ncaa_pts3 <- mutate(ncaa_pts2, seed_dif = win_seed - lose_seed)
head(ncaa_pts3)
```

```
##   win_seed lose_seed win_pts lose_pts win_school_ncaa pt_differ season round
## 1       11         6      94       90      Evansville         4   1989    64
## 2       14         6      75       63     Chattanooga        12   1997    32
## 3       14         3      73       70     Chattanooga         3   1997    64
## 4        4        13      61       39          Temple        22   1994    64
## 5        8         9      60       57          Temple         3   1985    64
## 6       10         7      80       63          Temple        17   1991    64
##   pts_dif seed_dif
## 1       4        5
## 2      12        8
## 3       3       11
## 4      22       -9
## 5       3       -1
## 6      17        3
```

Graphing the seed-based wins

The graph below is a smoothed histogram evaulating the winners in each game based on seed. Negative 15 means a 1st seed beat a sixteenth, negative numbers mean the higher seed won, and this occured roughly 70% of the time.

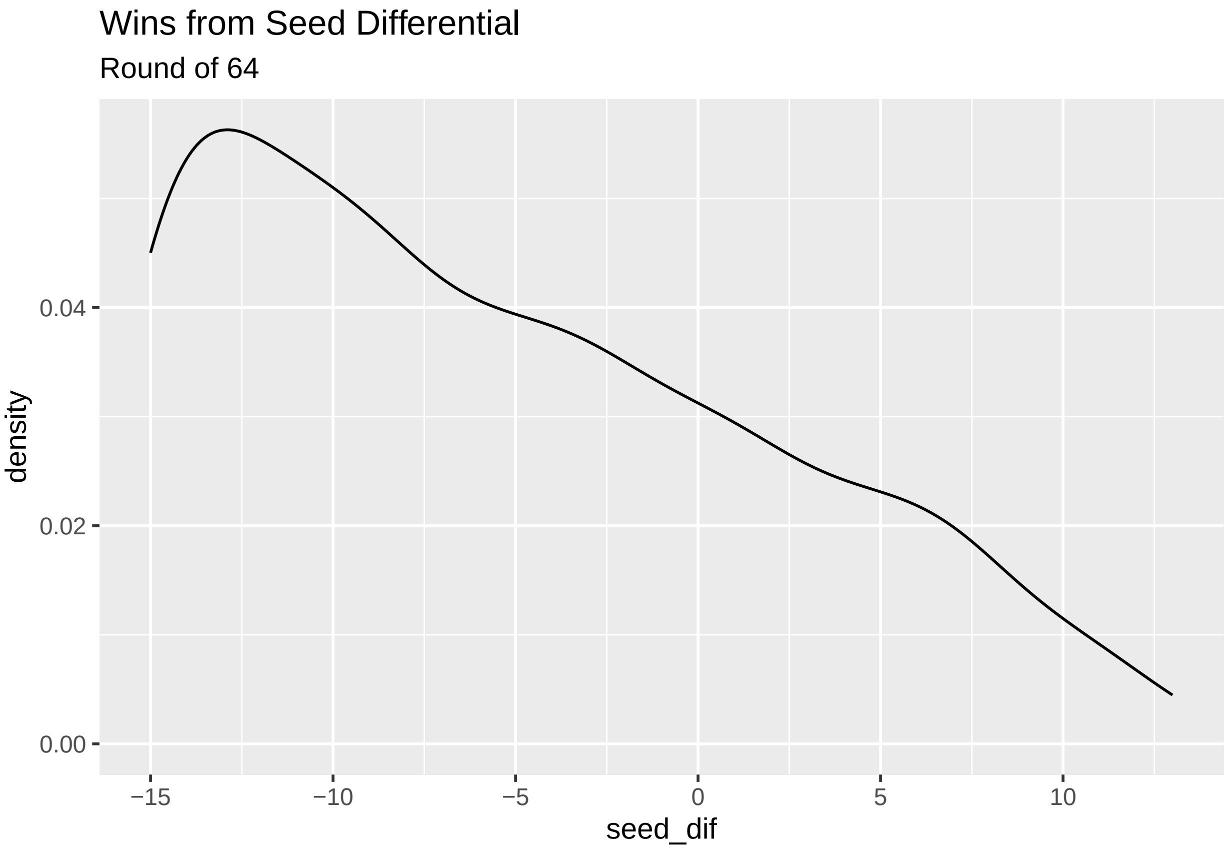
```
ggplot(ncaa_pts3, aes(seed_dif)) +geom_density()+
  labs(title = 'Wins from Seed Differential', subtitle = 'The full tournament')
```



Graphing the seed-based wins in the round of 64

This graph does the same as above but only with the round of 64. Even in the easiest round to do so, choosing the highest seed as the winner only gives you a 75% chance of guessing correctly.

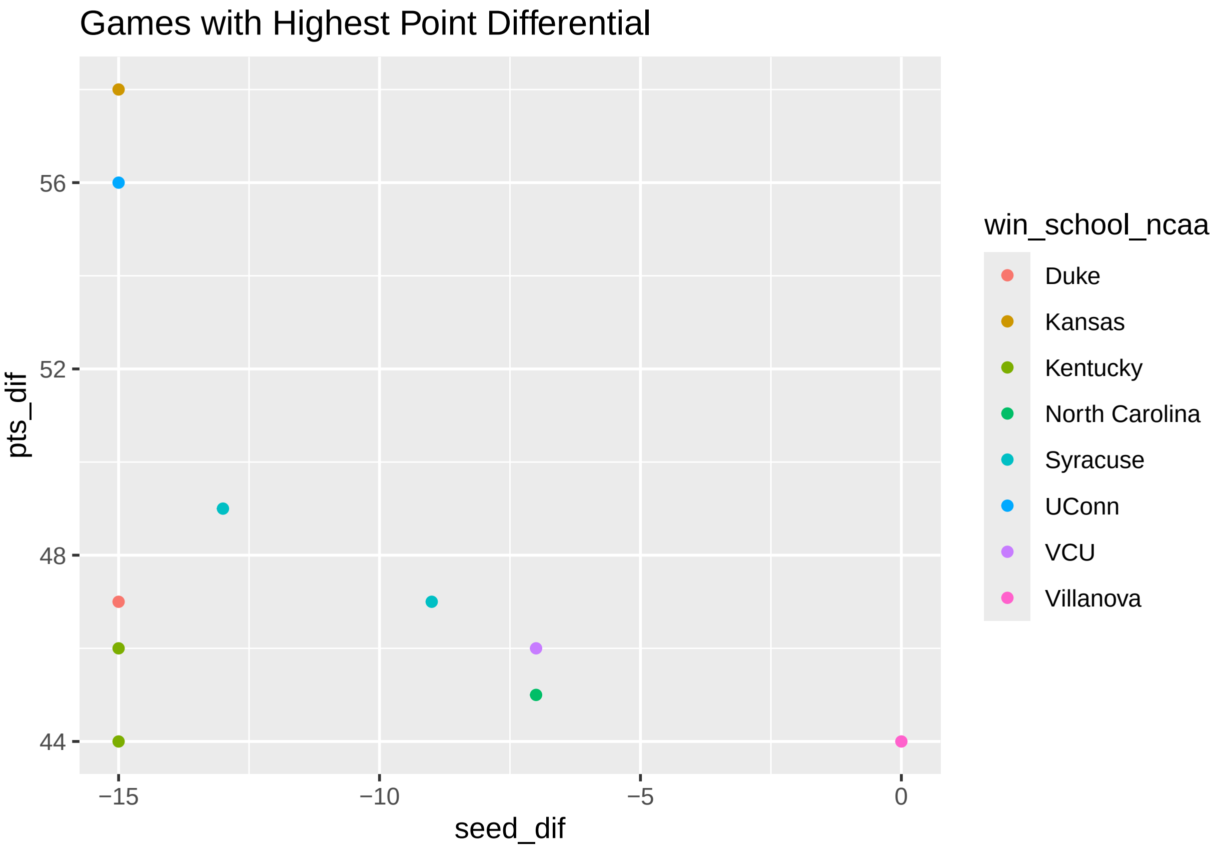
```
round_64 <- subset(ncaa_pts3, round == 64)
ggplot(round_64, aes(seed_dif)) +geom_density() +
  labs(title = 'Wins from Seed Differential', subtitle = 'Round of 64')
```



Team strength

What if you evaluate offensive and defensive strength based off of past performances? I have simulated this by using games within the tournament, but you could do so by evaluating a teams performance within their conference prior to the tournament. The graph below shows the winner of the 10 games with the highest point differential. Of these, only 2, Villanova and North Carolina, went on to win the tournament the same year as that high point differential game.

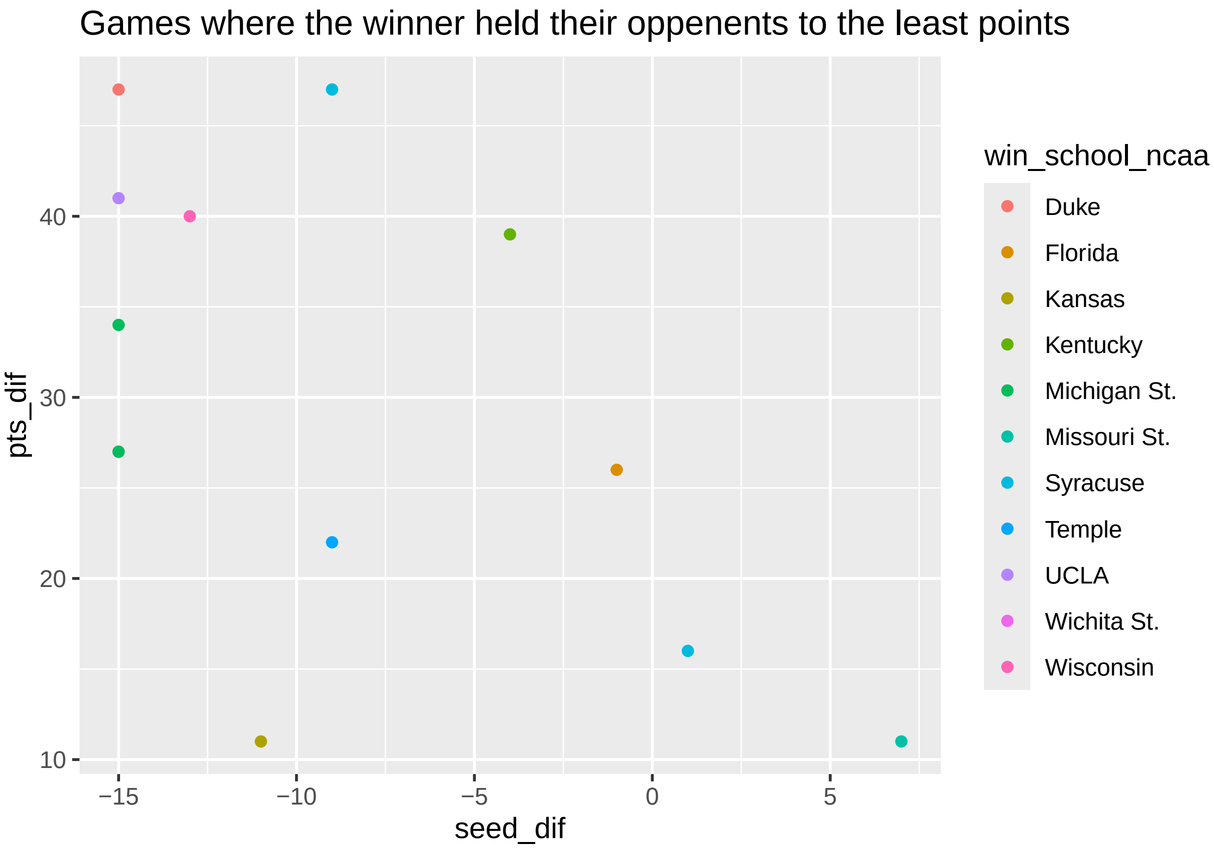
```
ncaa_pts3 %>%
  top_n(10, pts_dif) %>%
  ggplot() + aes(seed_dif, pts_dif, color = win_school_ncaa) + geom_point()+
  labs(title = 'Games with Highest Point Differential')
```



Defense wins Championships

What about defense alone? This graph shows the winners of the 10 games where the opposing team was held to the least points. Only one, Michigan St. in 2000, went on to be the champion.

```
ncaa_pts3 %>%
  top_n(-10, lose_pts) %>%
  ggplot() + aes(seed_dif, pts_dif, color = win_school_ncaa) + geom_point()+
  labs(title = 'Games where the winner held their oppenents to the least points')
```



Final Thoughts

So we've established that even with some data it's still difficult to guess the outome of the tournament, let alone each game. But what if you really knew your basketball? If you had an average 80% chance to guess each game correctly you would have a 1-in-476 million chance to get the entire bracket, or a little less than twice as difficult as winning the lottery. What if you really, really, *really*, knew basketball. If you had a 95% chance of predicting each game on average, then you'd be in good shape - a 3% chance of getting the entire bracket right, once every thirty years or so.