Project 2 - Tournament Unsupervised and Supervised Learning

William Sorg.74

2025-04-08

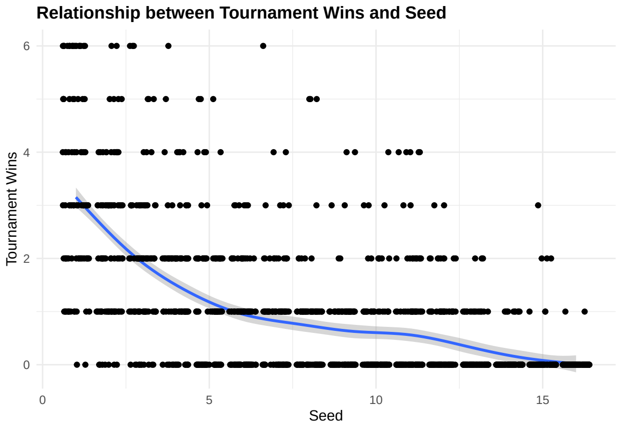
Data Selection

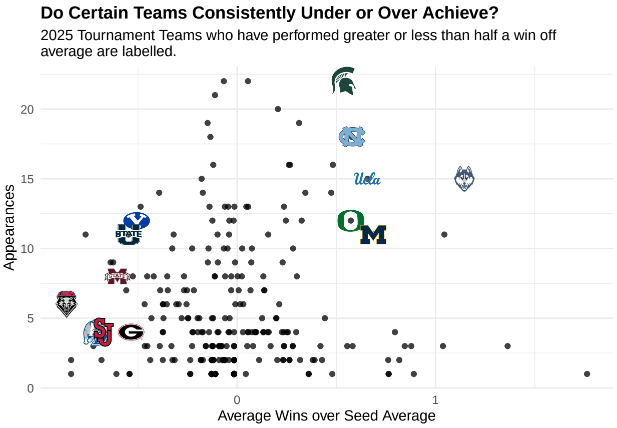
The data that I will be focusing my analysis on is a subset of the mm2002_2025.csv file. I removed all duplicate and redundant rows, then selected a subset of variables that I believe may be important. I also scraped the number of wins each team had in the tournament from sports-reference.com and joined this with the existing data.

```
## # A tibble: 6 x 40
##   Team      Season  Seed Region  wins Champ Finals FinalFour Conference AdjTempo
##   <chr>      <int> <int> <chr>  <dbl> <lgl> <lgl>  <lgl>     <chr>         <dbl>
## 1 Duke        2025     1 East       4 FALSE FALSE  FALSE     ACC            65.7
## 2 Alabama     2025     2 East       3 FALSE FALSE  FALSE     SEC            74.9
## 3 Wisconsin   2025     3 East       1 FALSE FALSE  FALSE     B10            67.6
## 4 Arizona     2025     4 East       2 FALSE FALSE  FALSE     B12            69.9
## 5 Oregon      2025     5 East       1 FALSE FALSE  FALSE     B10            67.6
## 6 BYU         2025     6 East       2 FALSE FALSE  FALSE     B12            67.2
## # i 30 more variables: AdjOE <dbl>, AdjDE <dbl>, AdjEM <dbl>, eFGPct <dbl>,
## #   TOPct <dbl>, ORPct <dbl>, FTRate <dbl>, OffFT <dbl>, Off2PtFG <dbl>,
## #   Off3PtFG <dbl>, FG3Pct <dbl>, FTPct <dbl>, DefFT <dbl>, Def2PtFG <dbl>,
## #   Def3PtFG <dbl>, BlockPct <dbl>, OppFG3Pct <dbl>, FG3Rate <dbl>,
## #   ARate <dbl>, AvgHeight <dbl>, CenterHeight <dbl>, Experience <dbl>,
## #   Bench <dbl>, PGPts <dbl>, SGPts <dbl>, SFPts <dbl>, PFPts <dbl>,
## #   CenterPts <dbl>, Net.Rating <dbl>, Active.Coaching.Length <int>
```

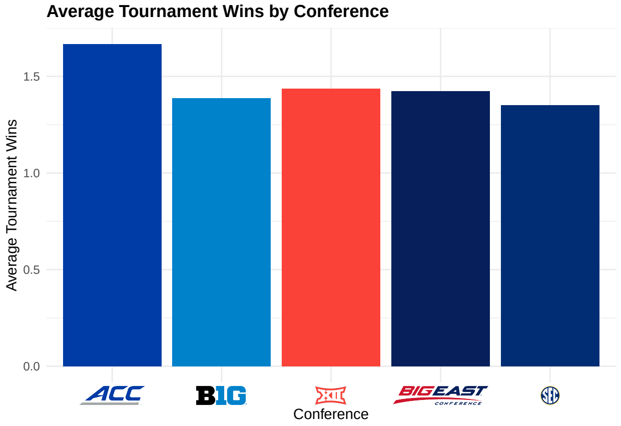
Exploratory Analysis

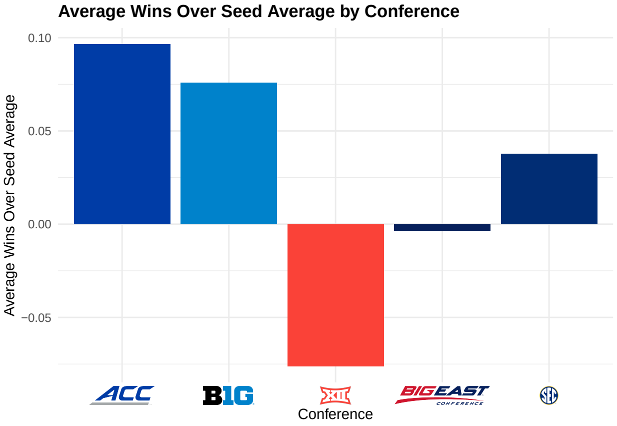
As expected, seed and tournament wins have a strong relationship. However, this relationship is the strongest for top seeds and tapers off for higher seeded teams. I was also curious if some teams are "built for March," so I looked at how teams performed compared to their seeds. While this isn't the most accurate metric since coaches matter more than the logo, it still gives some idea of how programs have performed in March Madness.



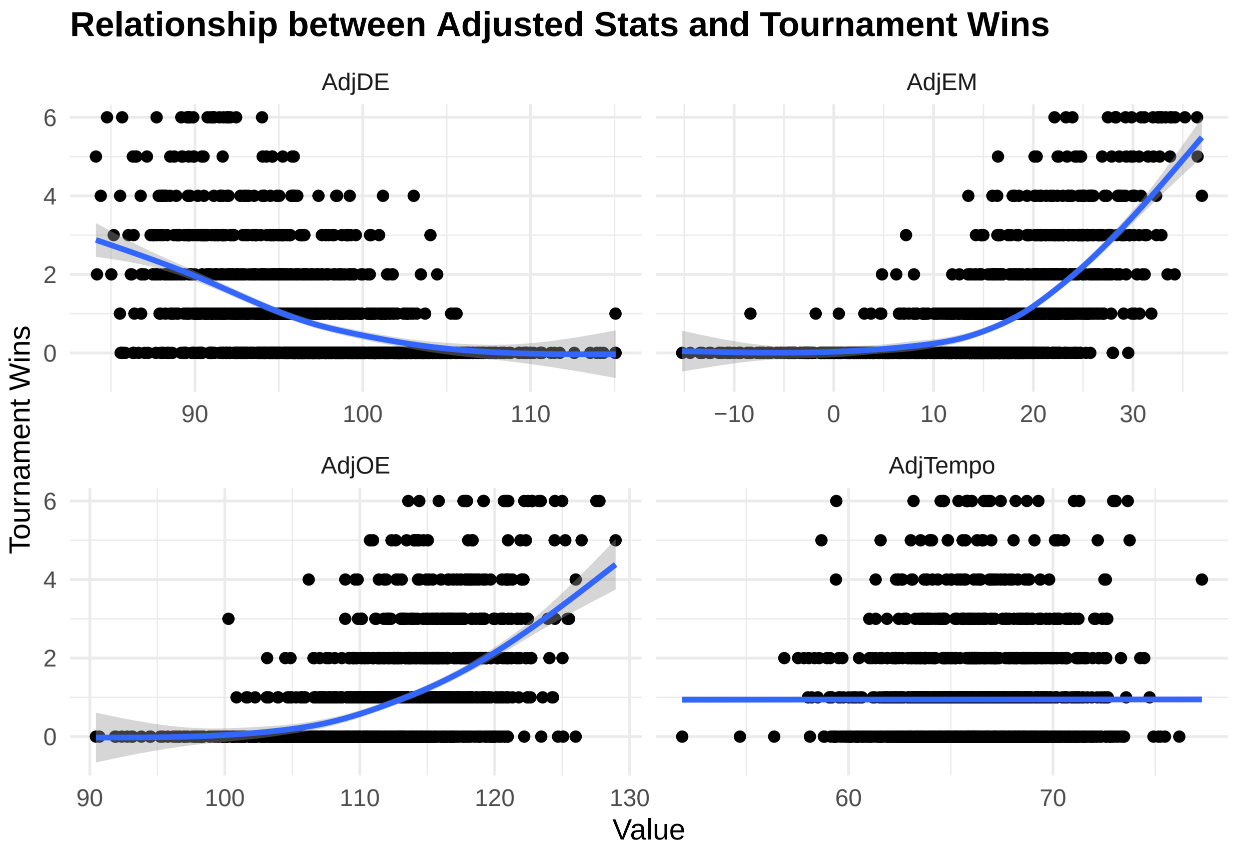


Based on the charts above, ACC teams tend to perform the best out of major conferences since 2002. There isn't a large difference in average tournament wins, but wins over seed average sees a wider range. I am surprised that the Big 12 under performs based on seed though.

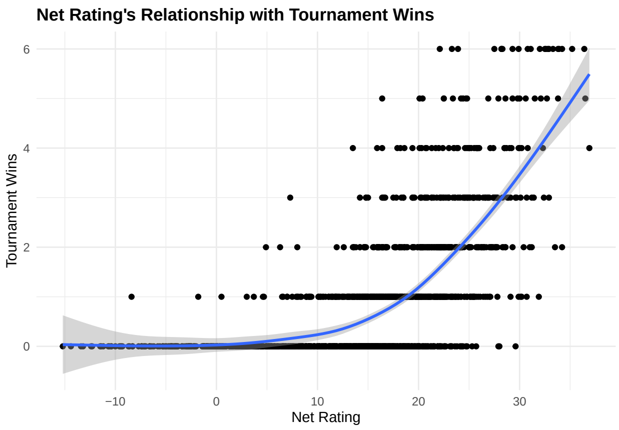


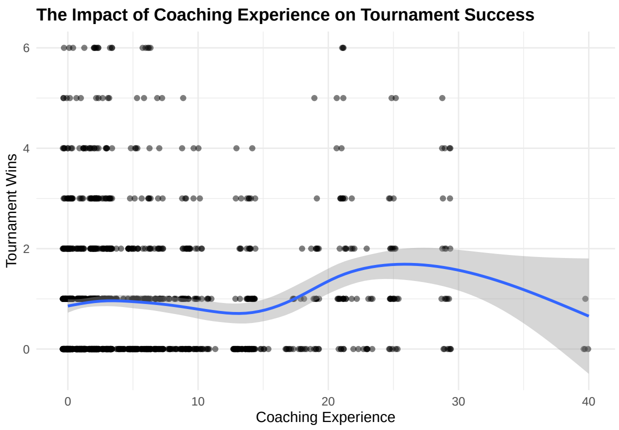


Adjusted Tempo doesn't seem to have much of an impact on tournament wins, but offensive and defensive efficiency, as well as their difference, do seem to have a significant relationship with tournament success.

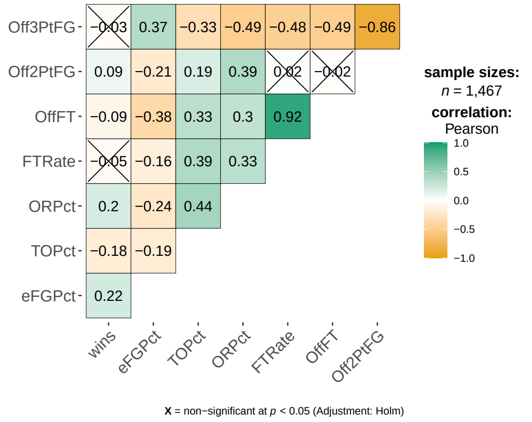


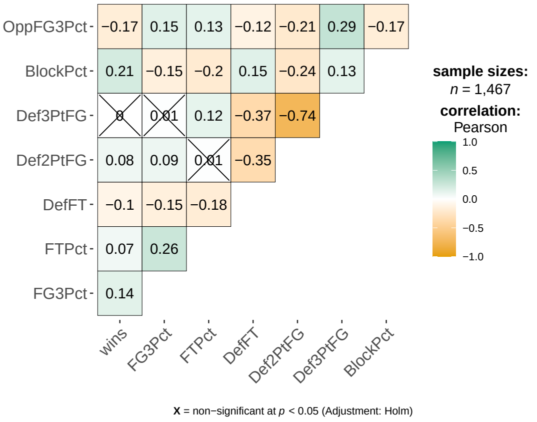
Net Rating appears to be a great predictor of tournament success while coaching experience does not. This is not surprising considering this years national championship game saw a 69 year old face a 39 year old.

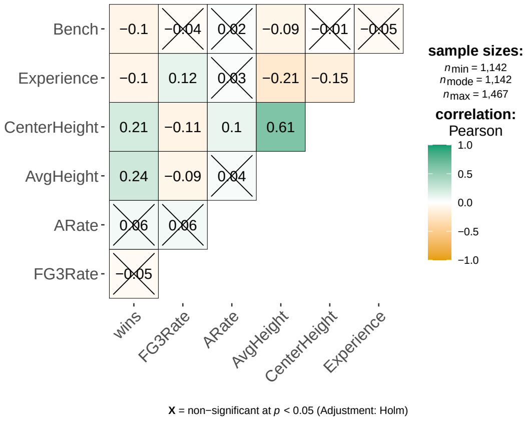




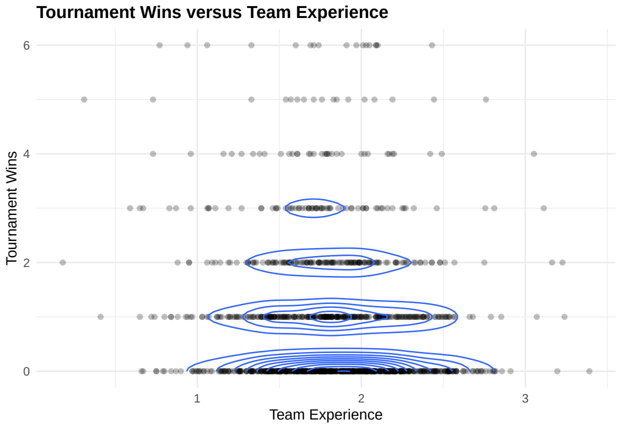
Given these correlograms, there appears to be no linear relationship between wins and many of these stats. However, offensive rebounding percentage, turnover percentage, effective field goal percentage, opponent 3-point field goal percentage, block percentage, center height, and average height all have a weak correlation with tournament wins.

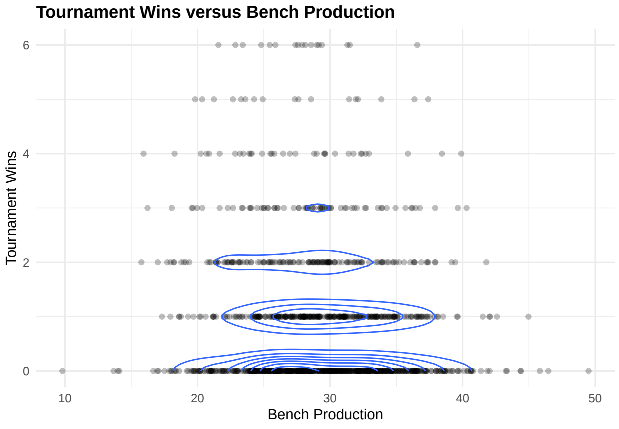




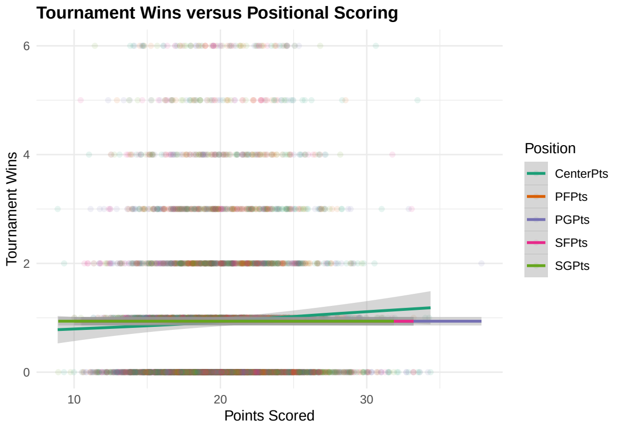


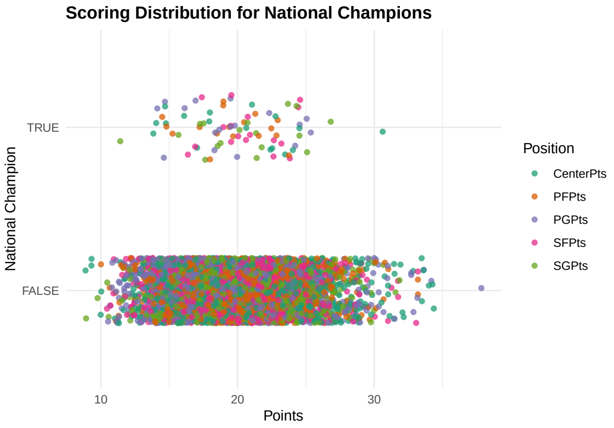
There isn't an extreme relationship between tournament wins and either team experience or bench production, but what is shown is that teams that are extreme in these categories don't tend to have success.





While positional scoring doesn't seem to have a relationship with tournament wins, all national champions appear to have somewhat balanced scoring.


Tournament Wins versus Positional Scoring


Scoring Distribution for National Champions