Project 2 - Tournament Unsupervised and Supervised Learning

William Sorg.74

2025-04-08

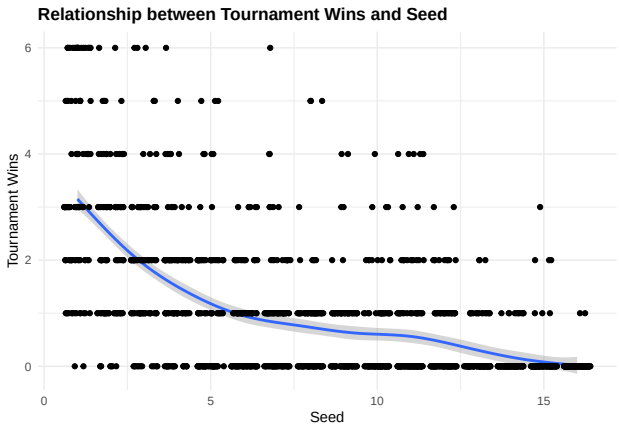
Data Selection

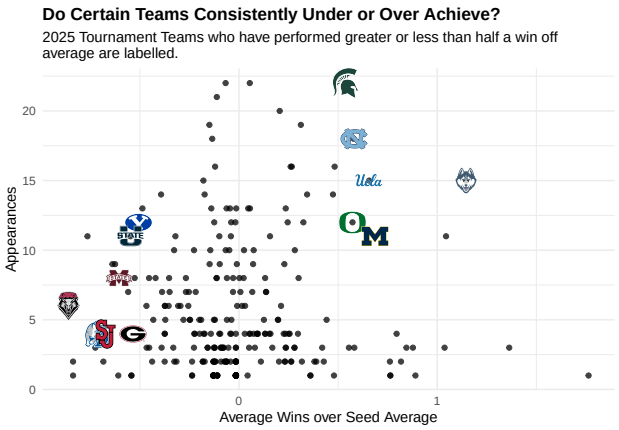
The data that I will be focusing my analysis on is a subset of the mm2002_2025.csv file. I removed all duplicate and redundant rows, then selected a subset of variables that I believe may be important. I also scraped the number of wins each team had in the tournament from sports-reference.com and joined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

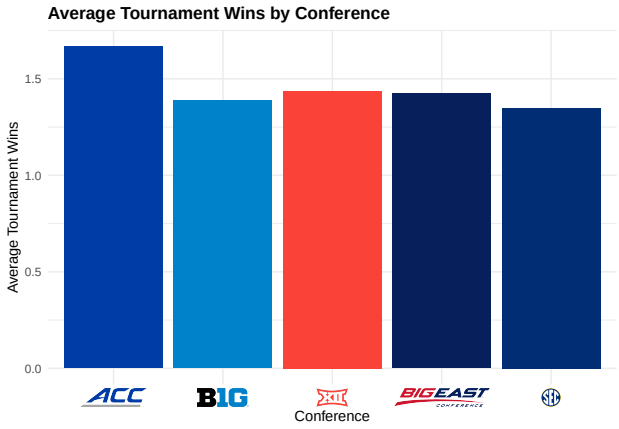
Exploratory Analysis

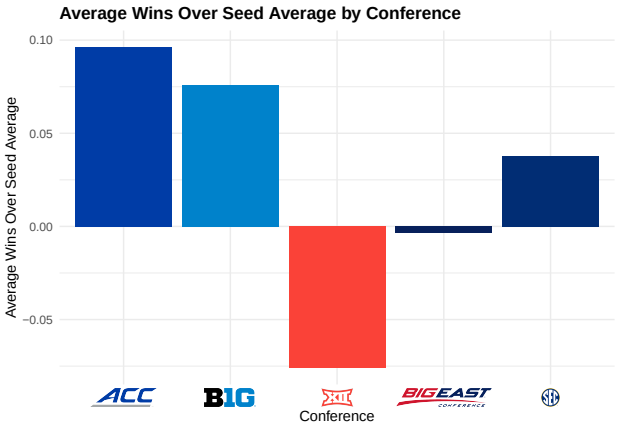
As expected, seed and tournament wins have a strong relationship. However, this relationship is the strongest for top seeds and tapers off for higher seeded teams. I was also curious if some teams are "built for March," so I looked at how teams performed compared to their seeds. While this isn't the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



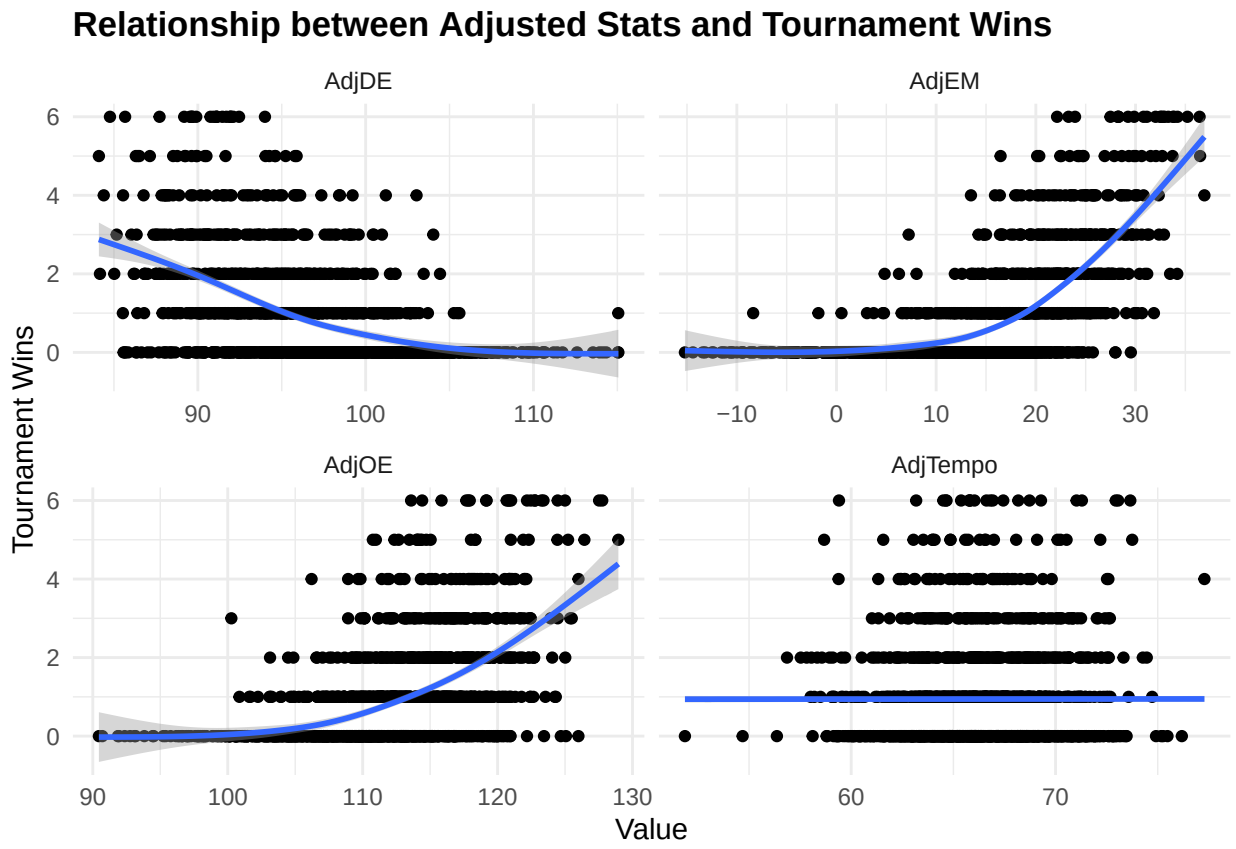


Based on the charts above, ACC teams tend to perform the best out of major conferences since 2002. There isn't a large difference in average tournament wins, but wins over seed average sees a wider range. I am surprised that the Big 12 under performs based on seed though.

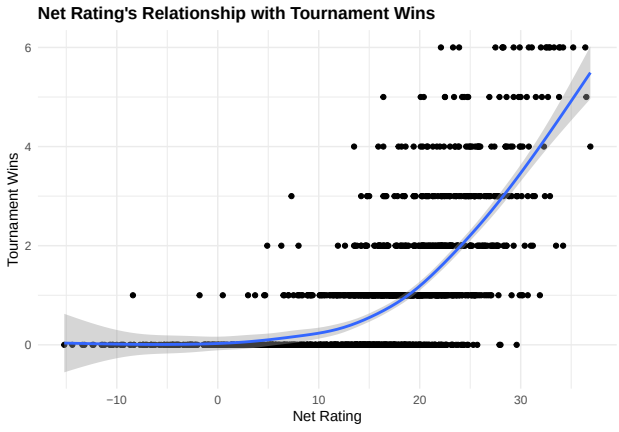


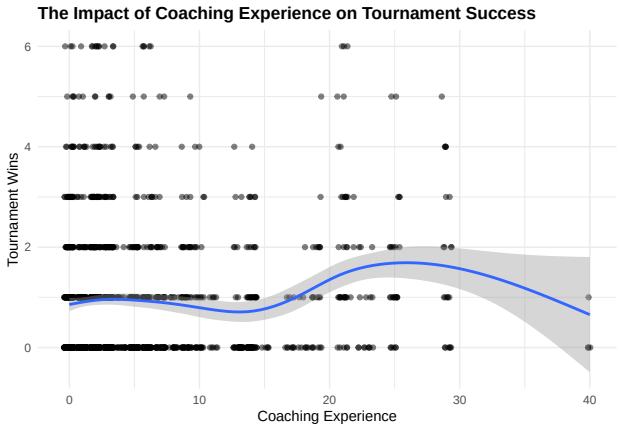


Adjusted Tempo doesn't seem to have much of an impact on tournament wins, but offensive and defensive efficiency, as well as their difference, do seem to have a significant relationship with tournament success.

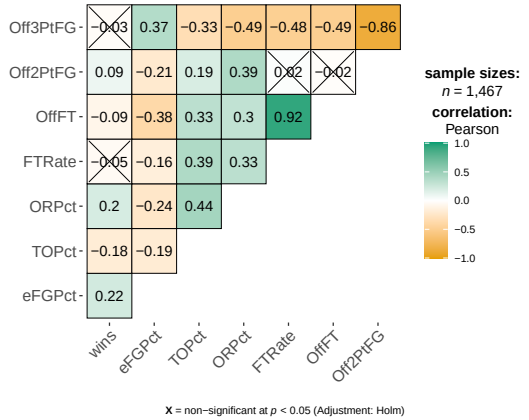


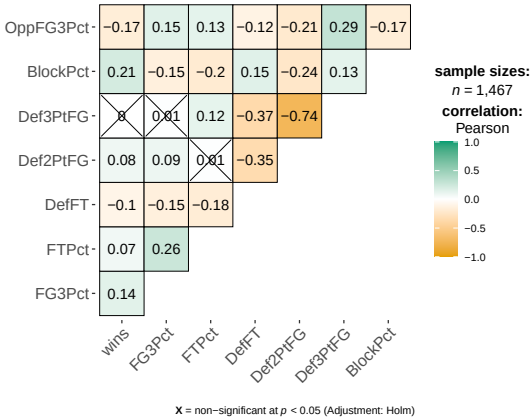
Net Rating appears to be a great predictor of tournament success while coaching experience does not. This is not surprising considering this years national championship game saw a 69 year old face a 39 year old.

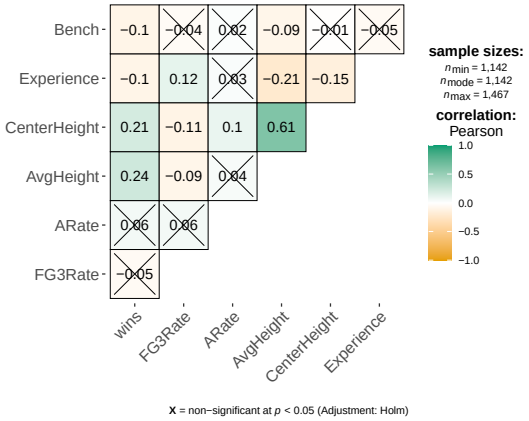




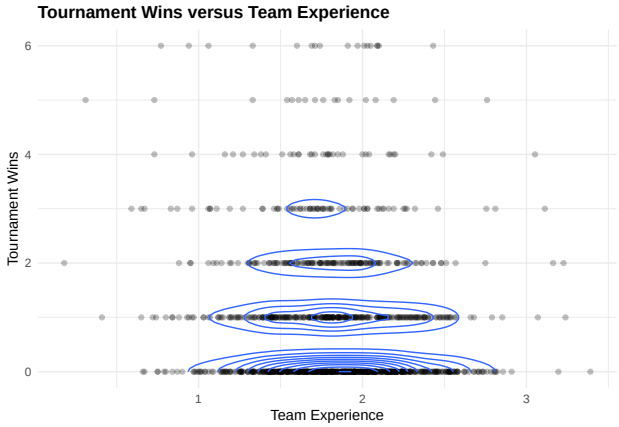
Given these correlograms, there appears to be no linear relationship between wins and many of these stats. However, offensive rebounding percentage, turnover percentage, effective field goal percentage, opponent 3-point field goal percentage, block percentage, center height, and average height all have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

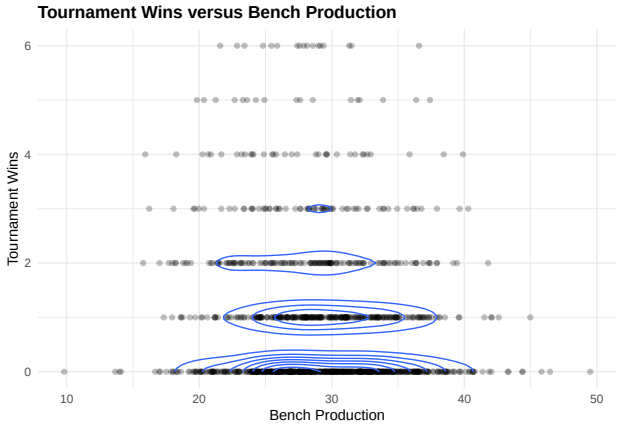




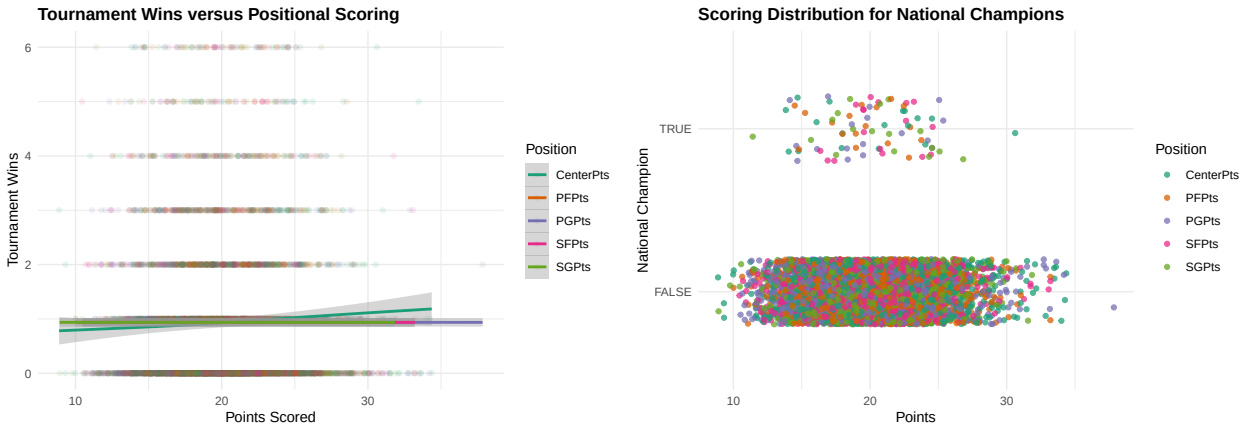


There isn't an extreme relationship between tournament wins and either team experience or bench production, but, what is shown, is that teams that are extreme in these categories don't tend to have success.





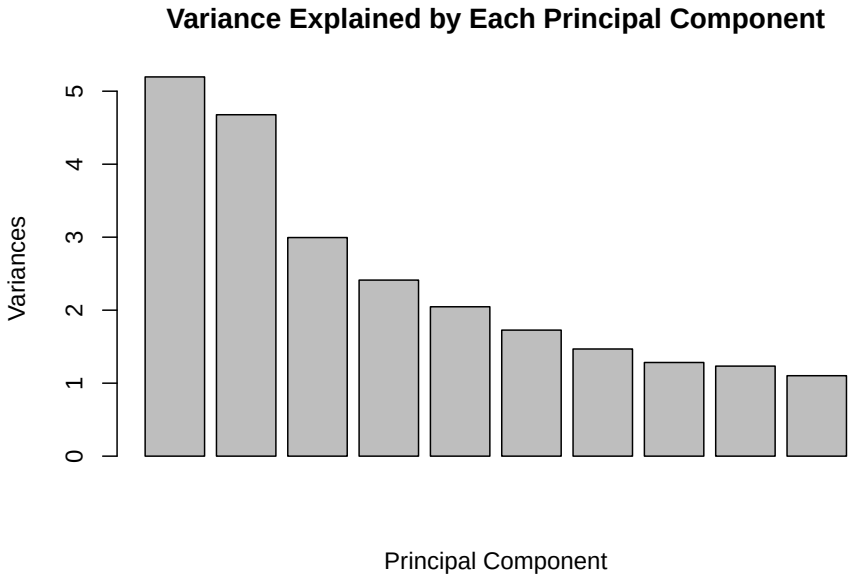
While positional scoring doesn't seem to have a relationship with tournament wins, all national champions appear to have somewhat balanced scoring.

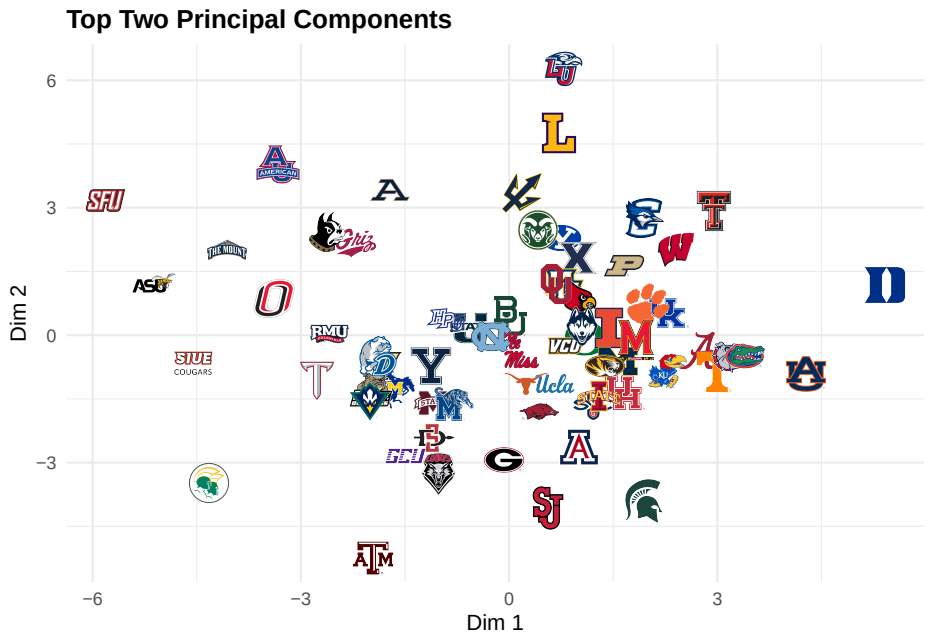


Unsupervised Learning

Principal Component Analysis

```
pcaout <- prcomp(pcadata[,10:40], scale. = TRUE, center = TRUE)
```

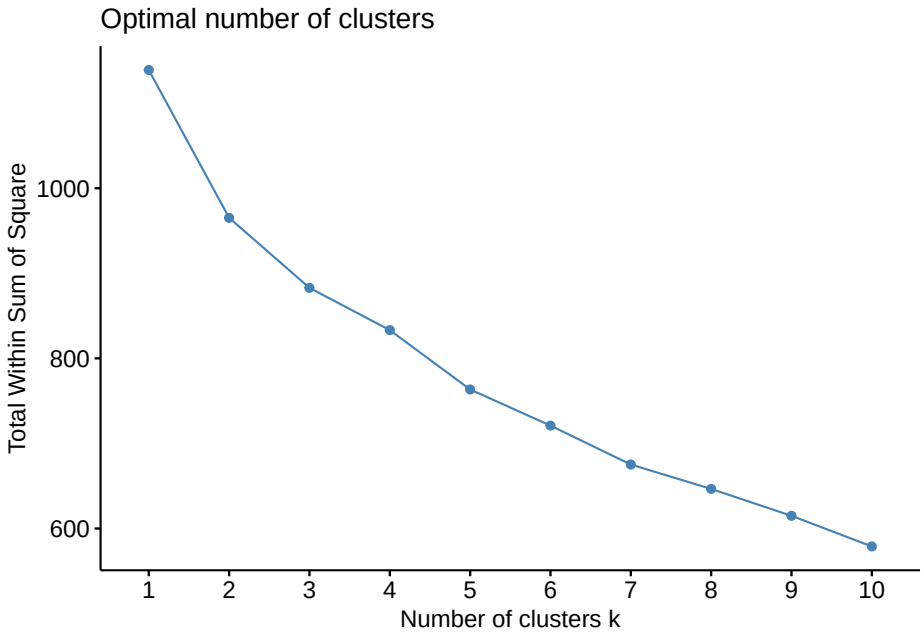




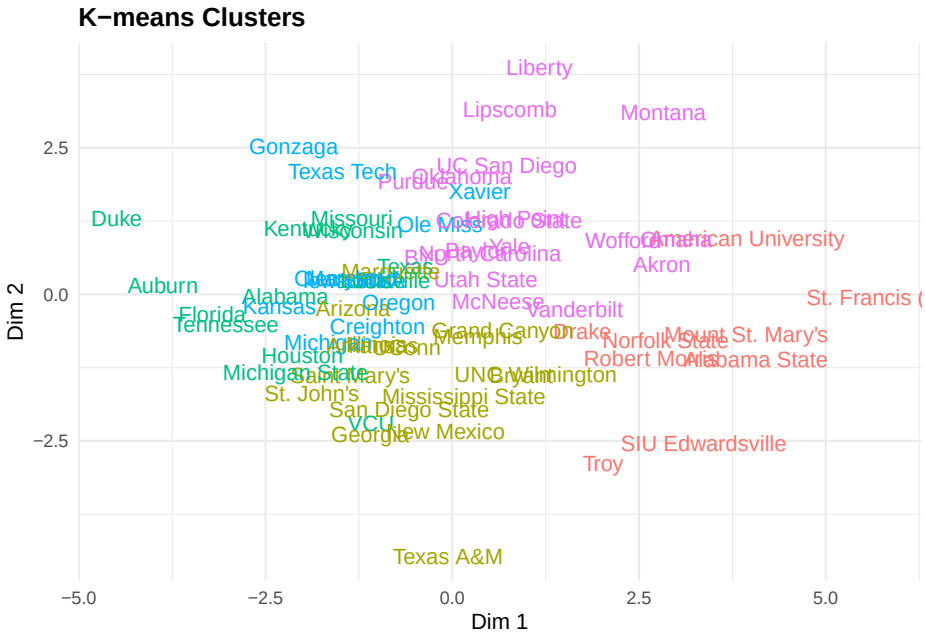
After projecting 31 variables down to two dimensions and retaining the maximum amount of variance, we get a plot that seems to have some trends. In dimension 1, better teams appear further to the right. 1-seeds Duke, Auburn, and Florida are the furthest right while high seeds are on the left. In dimension 2, it seems like the higher seed teams are further down. 2-seeds Saint John's and Michigan State are near the bottom while Liberty and Lipscomb are near the top. However, this dimension doesn't seem to have as noticeable of a trend. What sticks out the most is how alone Duke is compared to other teams.

K-means

```
clusterdata <- mmdata %>%
  filter(Season == 2025) %>%
  select(Team, AdjOE, AdjDE, Net.Rating, ORPct,
         TOPct, eFGPct, OppFG3Pct, BlockPct, CenterHeight, AvgHeight,
         Experience, Bench, PGPts, SGPts, SFPts, PFPts, CenterPts)

fviz_nbclust(scale(clusterdata[,2:18]), kmeans, method = 'wss', k.max = 10)
```




```
kmeansout <- kmeans(scale(clusterdata[,2:18]), centers = 5, nstart = 10)
```

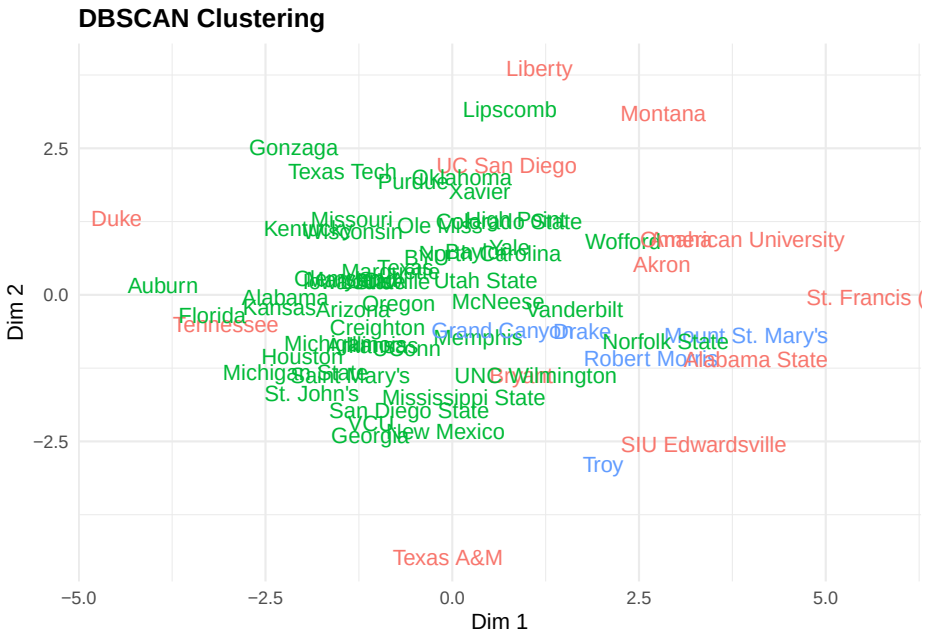


All teams were grouped into five clusters based on a smaller subset of the data. Looking at the five groups, the green cluster stands out with all four 1-seeds. Some surprises in this cluster are VCU and Missouri, who weren't highly seeded and lost in the first round. However, VCU was a trendy upset pick and Missouri had impressive regular season wins over Florida and Alabama, so I don't think these teams were necessarily bad. The red group also stands out. It has all four 16-seeds and is almost all bad teams, but Drake is in it. Drake was an 11-seed, but they won a game over aforementioned Missouri. They also gave Texas Tech a game in the second round. I think Drake is probably in this group because of the unique, slow playstyle.

The other three groups are much more of a mixed bag with some good and some bad. I think the biggest conclusion that can be drawn from being in one of these three groups is that they aren't in the elite or the trash cluster.

DBSCAN

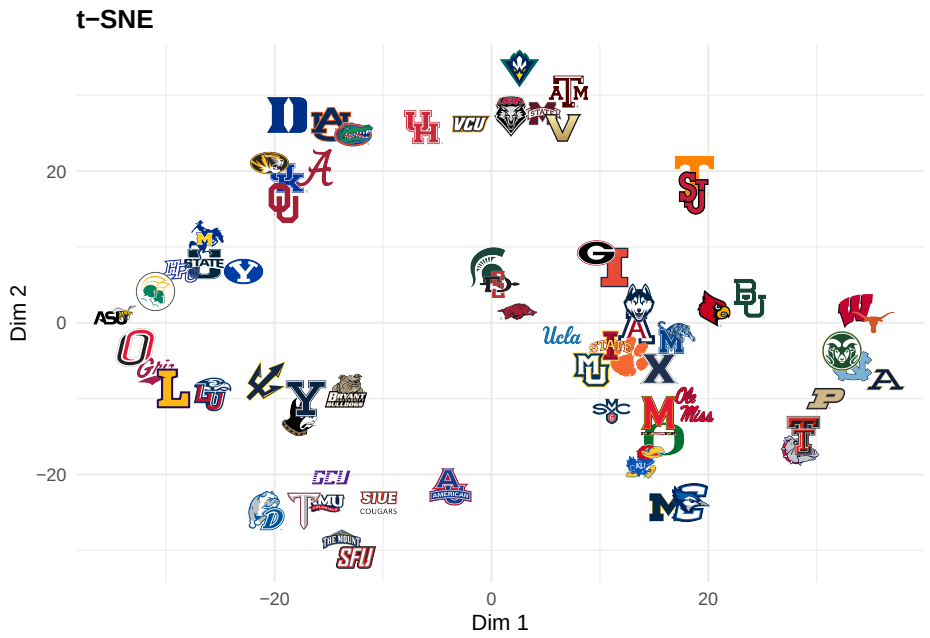
```
dbout <- dbscan(scale(clusterdata[,2:18]), minPts = 4, eps = 3.85)
```



DBSCAN clustered the majority of the teams in the tournament together, so I think it's best to focus on those not in the green cluster. The blue group doesn't seem like a great group of teams aside from Drake, who was oddly grouped in K-means as well. The outliers in the plot contain a wide variety of teams. On the right side, it's mostly bad teams. Near the center of Dimension 1 and on extremes of Dimension 2, you get three teams (A&M, Liberty, and UCSD) who all either won one game or nearly won one. Lastly, Duke and Tennessee were 1 and 2 seeds who reached the Final Four and Elite Eight, respectively. Overall, I don't think the DBSCAN clusters provide much information.

t-SNE

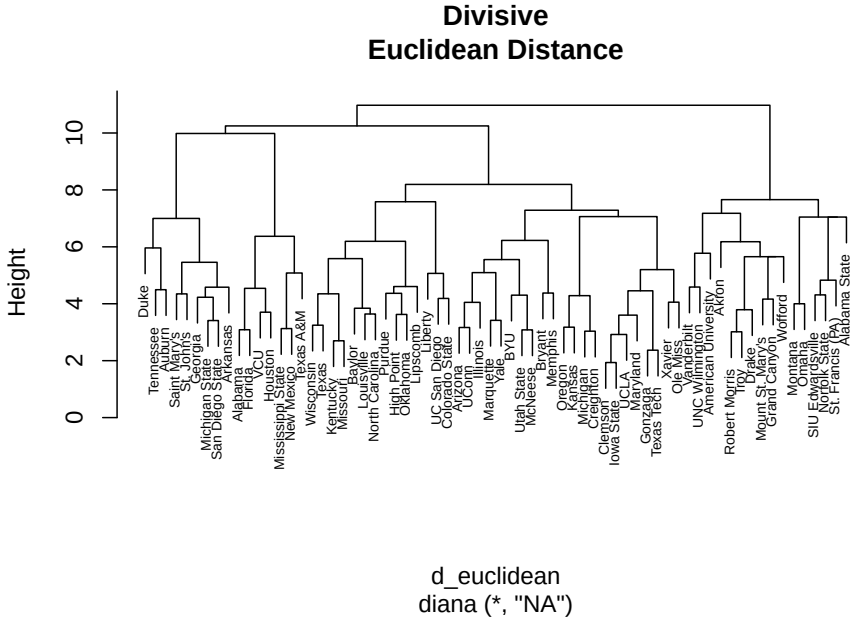
```
tsne_result <- Rtsne(scale(clusterdata[,2:18]),
                     perplexity = 4,
                     theta = 0.5,
                     dims = 2,
                     verbose = TRUE)
```

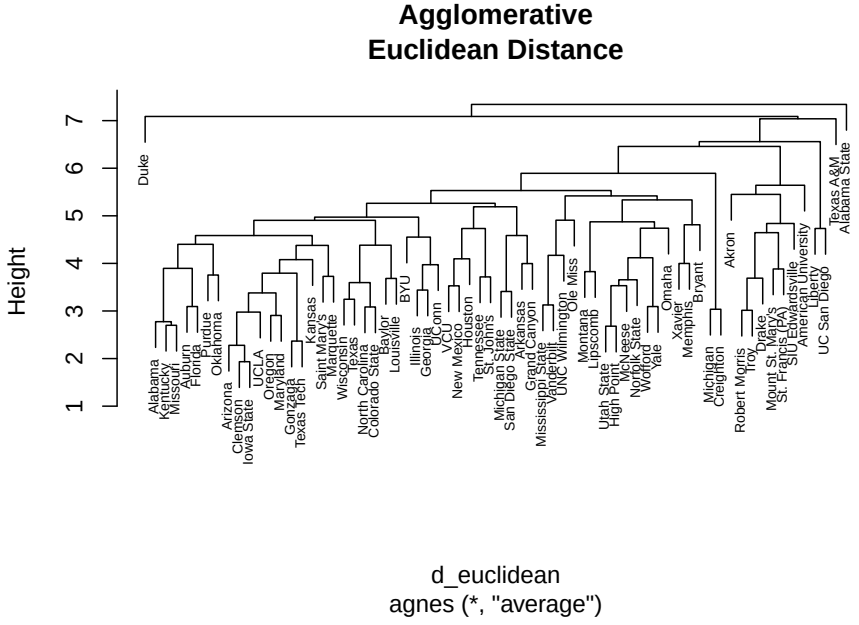


t-SNE gives another 2-dimensional representation of these teams. Some noticeable things are that all four 1-seeds are located closely together. VCU is close-by, as are a few SEC schools who had varying levels of success in the tournament. I am surprised by UNC Wilmington's location at the top of the plot. They did keep it close with Texas Tech in the first round though, so maybe they were underrated.
On the right side of the plot, they are all major conference teams aside from Akron. This is a bit surprising considering they got handled by Arizona in round 1. Although, the biggest surprise comes from BYU who's located near several sub par programs, despite reaching the Sweet 16.

Hierarchical Clustering

```
d_euclidean <- dist(scale(clusterdata[,2:18]), method = "euclidean")

diana_out <- diana(d_euclidean, diss = TRUE, metric = "euclidean")

agnes_out <- agnes(d_euclidean, diss = TRUE, metric = "euclidean")
```





These hierarchical clustering plots display many of the same trends from the other unsupervised methods. First, Duke is an outlier. Most of the plots have Duke at the extreme. Second, Drake's talent and tournamnent success are hard to distinguish using these methods. Drake is consistently grouped with the worst teams in the tournament despite being one of the stronger 11-seeds and a popular Cinderella pick. Lastly, VCU is a team that these methods suggest would be a strong tournament despite bowing out in the first round.

Supervised Learning

```
supervisedData <- mmdata %>%
  select(-AdjEM) %>%
  na.omit() %>%
  sample_frac()
trainData <- supervisedData %>%
  filter(Season != 2025)
testData <- supervisedData %>%
  filter(Season == 2025)
```

Elastic Net Regression

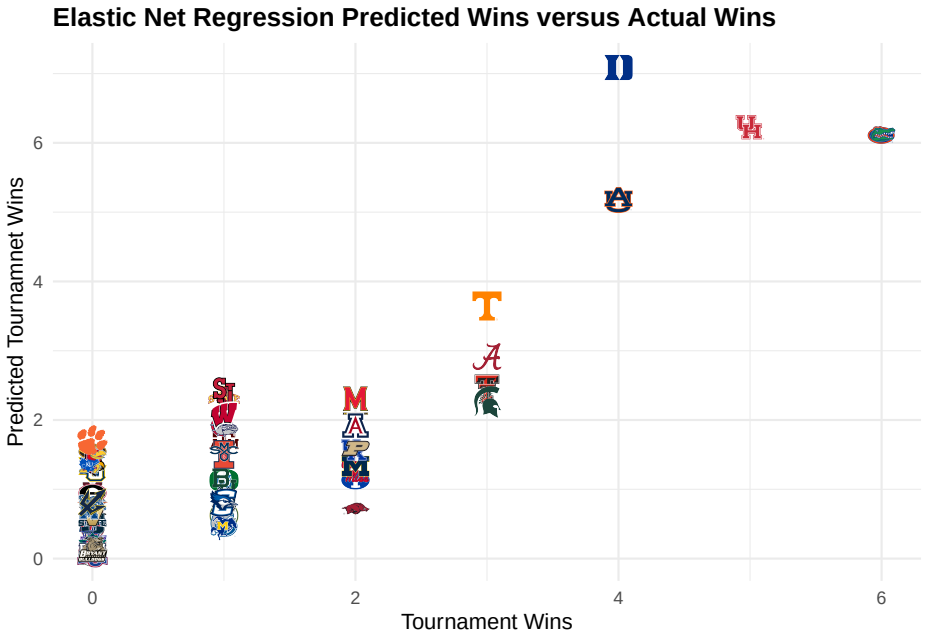
```
lambaOut <- cv.glmnet(as.matrix(trainData[,c(3,10:39)]), trainData$wins, nfolds = 10,
                      type.measure = 'mse', alpha = 0.5, family="poisson")

(lambda <- lambaOut$lambda.min)
```

```
## [1] 0.07411082
```

```
netOut <- glmnet(as.matrix(trainData[,c(3,10:39)]), trainData$wins, nfolds = 10,
                 type.measure = 'mse', alpha = 0.5, lambda = lambda, family="poisson" )

netOut$beta
```

```
## 31 x 1 sparse Matrix of class "dgCMatrix"
##                                    s0
## Seed                   -0.031354113
## AdjTempo                .
## AdjOE                   0.038987834
## AdjDE                  -0.043282923
## eFGPct                  .
## TOPct                   .
## ORPct                   .
## FTRate                  .
## OffFT                   .
## Off2PtFG                .
## Off3PtFG                .
## FG3Pct                  .
## FTPct                   .
## DefFT                   .
## Def2PtFG                .
## Def3PtFG                .
## BlockPct                .
## OppFG3Pct               .
## FG3Rate                 .
## ARate                  -0.007945562
## AvgHeight               .
## CenterHeight            .
```

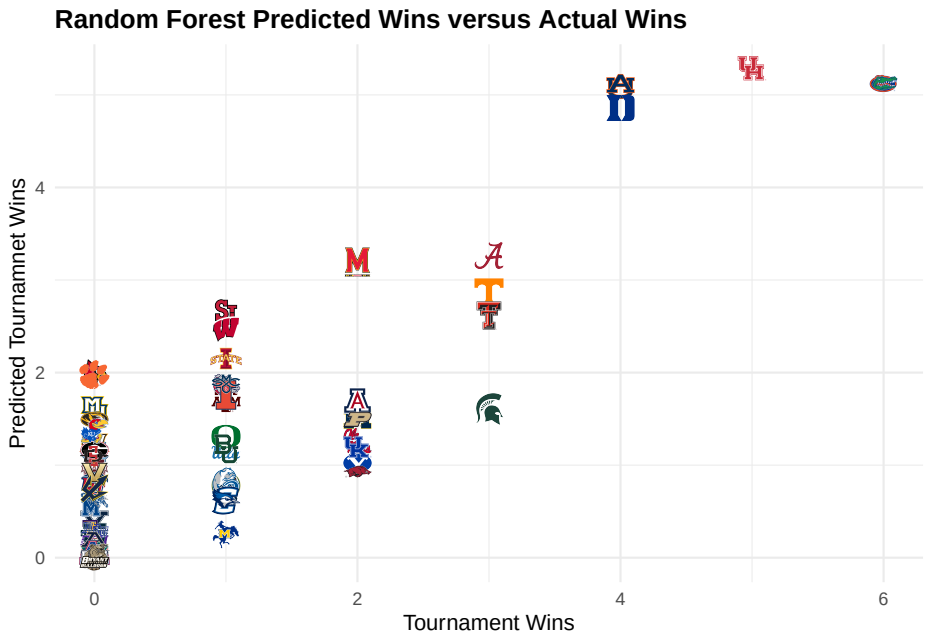
```
## Experience                .
## Bench                    -0.008324444
## PGPts                     .
## SGPts                     .
## SFPts                     .
## PFPts                     .
## CenterPts                 .
## Net.Rating                0.057666520
## Active.Coaching.Length    0.001149497
```



This elastic net regression combines LASSO and Ridge to penalize the model for using uninformative variables in the Poisson regression. This method makes it clear who the top four teams are, but it believes Duke is the best. It also does a nice job of projecting Elite Eight teams, since the four teams who lost in the Elite Eight were expected to get around 3 wins in the tournament.
I will use the β values from this regression to inform variable choices for the remaining supervised learning methods.

Random Forest

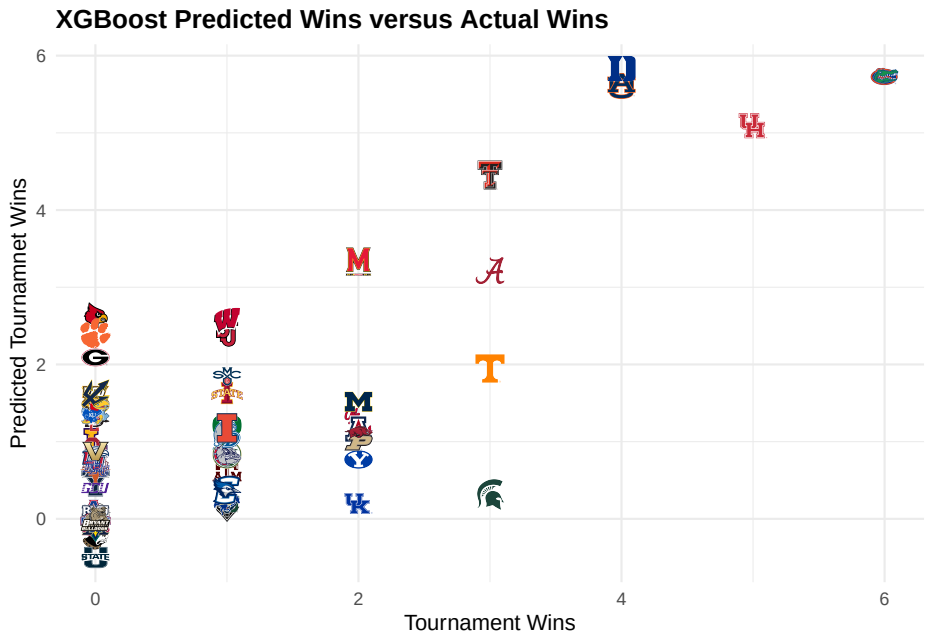
```
rfout <- randomForest(wins ~ Seed + AdjOE + AdjDE + ARate + Bench + Net.Rating,
                      data = trainData, mtry = 3)
```



Applying a random forest model to predict tournament wins given six relevant variables. The random forest model helps to avoid overfitting our training data. Once again, the top four teams, by seed and result, are correctly predicted. However, this time the margin between them is much smaller and Houston is expected to win the most games. Another notable difference is that Michigan State is expected to win much less games and Maryland is expected to win much more. This method has Maryland as the sixth best team in the tournament. Also, Clemson and Louisville were expected to win two games (not taking into account their matchups), but didn't even win one.

XG Boost

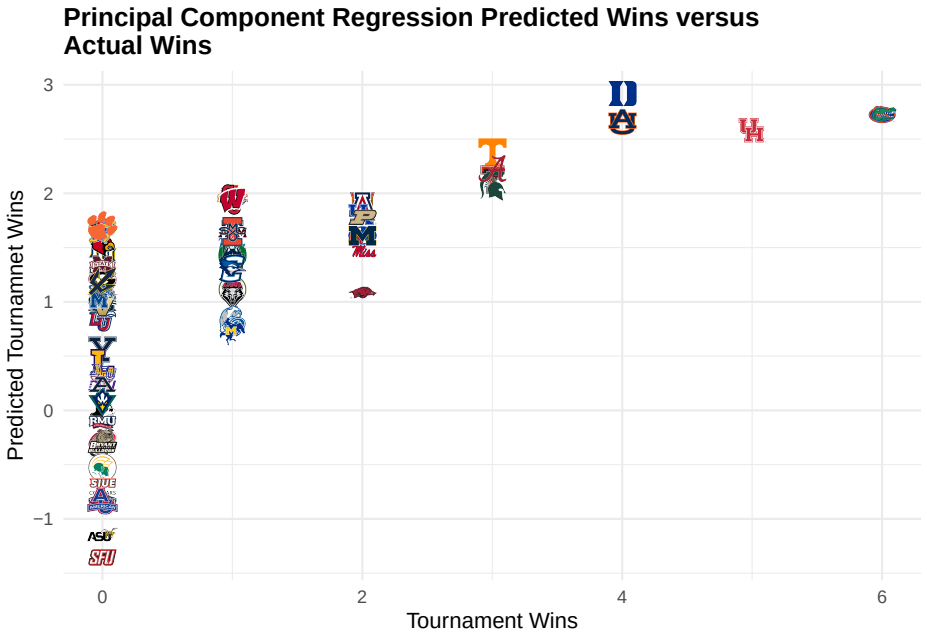
```
xgMatrix <- xgb.DMatrix(as.matrix(trainData[, c(3,11,12, 28, 32, 38)]),
            label = trainData$wins)

xgbOut <- xgboost(data = xgMatrix, nrounds = 200)
```



Like the previous two models, an XGBoost model has the four 1-seeds as the best four teams. This method has Duke as the best, followed by Florida, Auburn, and Houston. The big changes occur outside of these four teams. The model really likes Texas Tech and would've expected them to make the Final Four in a typical year. The XGBoost model also dislikes Michigan State, even more than the random forest model. This doesn't even expect Michigan State to win a game in a typical year. Similarly, this model doesn't believe in Kentucky. Lastly, Louisville, Clemson, and Georgia were expected to have more success.

Principal Component Regression

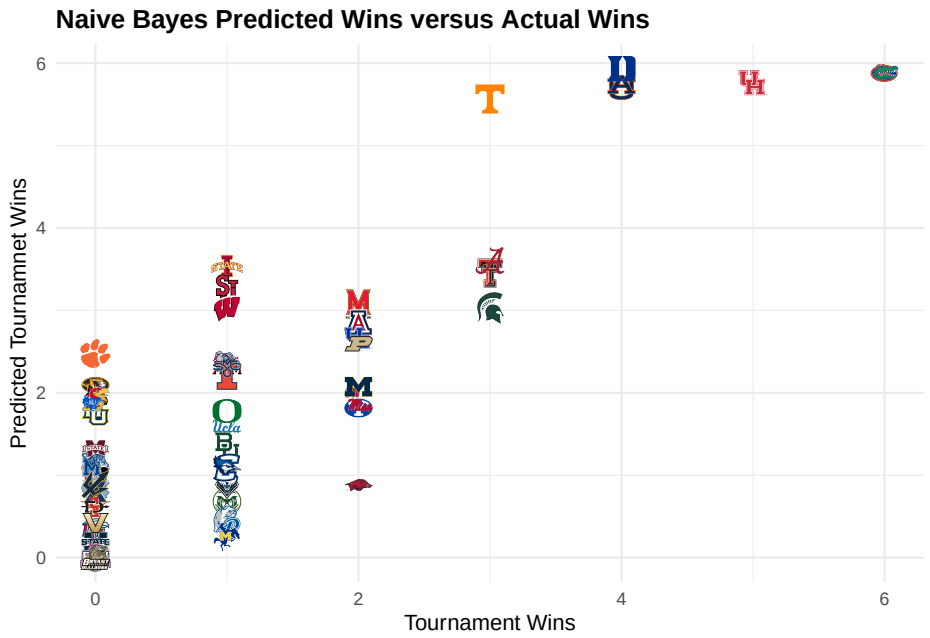
```
pcm1 <- pcr(wins ~ Seed + AdjOE + AdjDE + ARate + Bench + Net.Rating,
            data = trainData)
```



Principal Component Regression returns much lower expected win totals across the board, but the patterns remain the same. Duke is considered the top team along with the other 1-seeds. This method does return the correct Elite Eight teams, with Tennessee, Alabama, Texas Tech, and Michigan State as the next four best. After that, the rest of the field is more clumped together. Arkansas is the clear overachiever in this graph, while Clemson and Kansas underachieved.

Naive Bayes

```
nbout <- naiveBayes(wins ~ Seed + AdjOE + AdjDE + ARate + Bench + Net.Rating,
                    data = trainData, type = 'raw')
```



Like every other supervised method, the top four teams are the four 1-seeds when using Naive Bayes. However, Tennessee is also included in this elite tier. Again, Arkansas stands out as the Sweet 16 overachiever, while Iowa State, St. John's, and Wisconsin were expected to do much better than the Round of 32. Similar to the previous results, Clemson was considered the best team to lose in the first round. Missouri, Louisville, Kansas, and Marquette were also disappointments based on this model.

Analysis and Findings

After using several different supervised and unsupervised methods to predict the results of the 2025 Men's NCAA Tournament, the viability of the results can now be analyzed. When analyzing results, emphasis will be placed on picks further into the tournament that generate more points in bracket challenges. For this reason, we will begin by looking at the Final Four and the National Championship.

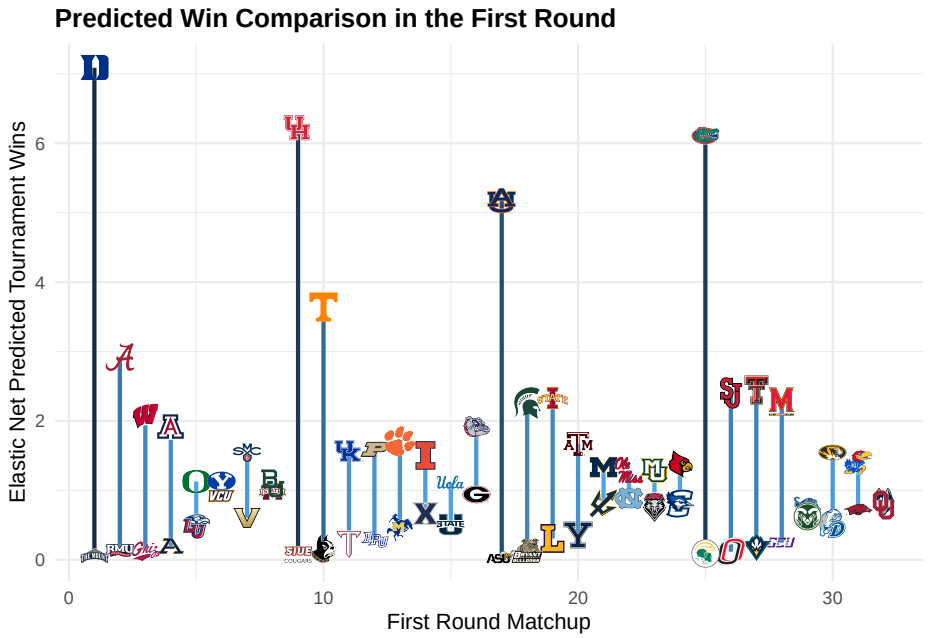
The Final Four consisted of Auburn, Florida, Duke, and Houston, all of which were 1-seeds. This was the first time since 2008 that all four No. 1 seeds reached the Final Four, but given the data, it's not entirely surprising. In supervised models, these four teams consistently graded out a tier above the rest of the field. Across all methods, it was clear that these were the best four teams in the country. A negative impact of this, though, is that it makes picking the National Championship much more difficult.

In the Duke-Houston matchup, I think the clear choice was Duke. They were better than Houston in 4 out of 5 supervised methods and in unsupervised methods, Houston was sometimes clustered with lesser teams. Even though Duke ultimately lost to Houston, I believe Duke was the logical pick. The Auburn-Florida matchup is a more difficult decision. They were consistently clustered together and graded out similarly in 4 out of 5 supervised methods. Ultimately, I think the decision would have to come down to each team's momentum entering the tournament. Florida was coming off an SEC Championship while Auburn had lost 3 of 4. Because of this, I think Florida is the correct choice.

For the national champion, I still think Duke was the correct choice. They're predicted to have more wins

than Florida in almost every model and seemed like they were a step above Florida. Obviously, Florida won the championship, but I believe this Duke team was one of the best teams from the past decade.

Now looking at the Elite Eight, there are four spots remaining considering we already have the four 1-seeds penciled in. Based on the plots, I think Tennessee, Alabama, and Texas Tech seem like strong choices. As for this final spot, Michigan State is loved by many of the models and Iowa State suffered some major injuries heading into the tournament. Because of this, this a spot where I think picking an upset is a fun idea.



Choosing upsets is a critical part of filling out a bracket. While having a chalky bracket may seem like (and probably is) the smart choice, it isn't as fun. Considering the expected wins above, we identify BYU-VCU, Michigan-UCSD, Ole Miss-UNC, and Memphis-CSU as potential upset spots (excluding 7-10 and 8-9 matchups). Of these, only one occurred, and the other three saw the team on upset watch win two games. Unfortunately, I initially picked all four of these games incorrectly, but I don't feel bad about it because they were among the better options in a tournament with very few upsets. The biggest upset of the first round was McNeese State over Clemson, which surprised me greatly at the time and still does. McNeese State was a popular pick due to coach Will Wade and their viral team manager. However, I selected Clemson to reach the Sweet 16 based on their performance last year and the buzz surrounding a possible upset. This turned out to be incorrect, and McNeese State exemplified what's special about March Madness: its unpredictability.

Creating an NCAA Tournament bracket involves numerous challenges and decisions. Utilizing various statistical and machine learning algorithms can help simplify these decisions. However, there are still several limitations. For one, it can be challenging to consider the impact of injuries. This could involve losing a player before the tournament (Kentucky or Iowa State) or dealing with injuries during the tournament (Auburn). Another limitation is measuring how "hot" a team or player is. This year, Florida entered the tournament playing their best basketball and ended the season by cutting down the nets. Additionally, nearly every year, we witness a player who steals the spotlight. Walter Clayton Jr. accomplished this feat this year, but predicting who will be the next March legend can be tricky. The most significant challenge, however, is that we are dealing with collegiate athletes. The performance of these athletes naturally fluctuates, and March Madness only heightens that variability. Ultimately, while statistics can help ease the challenge of

filling out your bracket, it's crucial to not take your picks too seriously and to consider selecting a few upsets. Life is too short for chalky brackets.