Does Defense Win Championships?

Anlian Krishnamurthy, Dooley Kim, Nick Strezo, Zach Liptzin

Introduction

The NCAA basketball tournament, better known as March Madness, is one of the most anticipated sporting events in American society. The tournament, famous for its thrilling games, unpredictable upsets, and the drama of the win or go home format brings the attention of millions of fans every year. The tournament brings in millions of dollars in revenue every year, while also giving schools a valuable platform to advertise and showcase their institution. With a lot on the line, people often debate over what the best predictor of success is. Is it an unstoppable offense or a lockdown defense? By analyzing which type of stats are better predictors of performance, we can better understand what it takes to succeed in one of the most competitive environments in sports.

We used a dataset full of college basketball statistics from the last ten years. This paper will analyze numerous statistics in comparison to a postseason score we have assigned based on success in the NCAA tournament. A team's conference is also factored in as a categorical variable. Overall, our main question is whether offense or defense is more of a factor in postseason success.

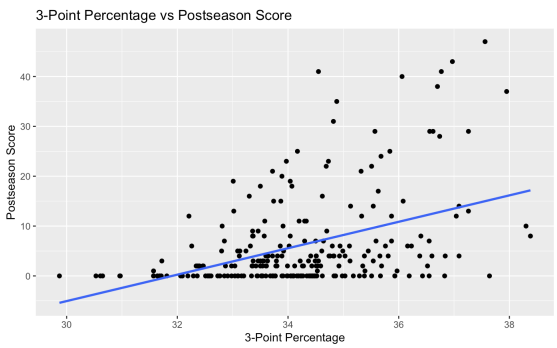
Our research question is important because many media outlets claim that defense wins championships and we want to run the numbers to figure out if this is true or not. Our research aims to answer the question of whether the role of defensive performance or offensive performance has more effect in determining a team's overall winning percentage. There are obviously a lot of variables that contribute to a team's success such as coaching, funding, and recruiting. However, examining specific offensive and defensive stats such as free throw rate and offensive rebounds allows us to provide a clearer understanding of what contributes more to winning.

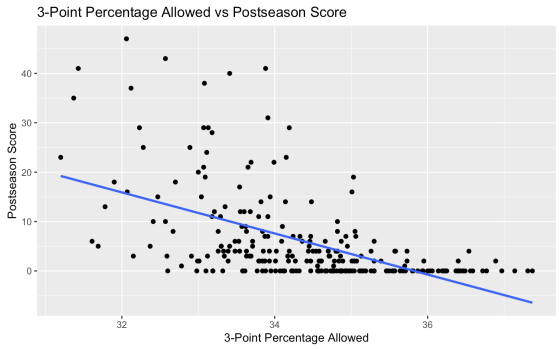
Data and Exploratory Data Analysis

We used Effective Field-Goal Percentage, Effective Field-Goal Percentage allowed, 3-Point Percentage, 3-Point Percentage Allowed, Turnovers Per Game, Turnovers Forced Per Game, Free Throw Rate, and Free Throws Allowed as predictor variables and postseason score as the outcome variable. We also used conference as a confounder, specifically separating the "Power 6" conferences from the rest. We

obtained the data through a dataset from Kaggle that includes stats from all D1 college basketball teams. We created the postseason score with the following system: 1 point for making the first four, 2 points for making the round of 64, 3 points for making the round of 32, 4 points for making the sweet 16, 5 points for making the elite eight, 6 points for making the final four, 7 points for making the final, and 8 points for winning the championship.

3-Pointers





These two graphs are 3-point percentage allowed and 3-point percentage plotted against postseason score. As you can see, shooting a higher percentage from the 3-point line leads to a higher postseason score, and allowing your opponents a higher 3-point percentage leads to a lower postseason score. To figure out which metric has a greater impact on postseason score, we conducted the following linear regressions:

$$PostseasonScore = B0 + B1*3PointPercentage$$
$$PostseasonScore = B0 + B1*3PointPercentageAllowed$$

After running the regressions, I got 2.65 for B1 in the first model and -4.16 for B1 in the second model. Both had p-values that were very very small, meaning we can reject the null hypothesis that these variables have no effect, and can now say the two values are statistically significant. Since percentages are generally in the 30s range, the intercepts are meaningless because they give what's happening when a team is shooting or allowing 0% from the 3-point range. Using just these variables, it appears that 3-point percentage allowed has a bigger impact on postseason score. This would support the claim that defense has a greater impact on postseason success.

FG Percentage

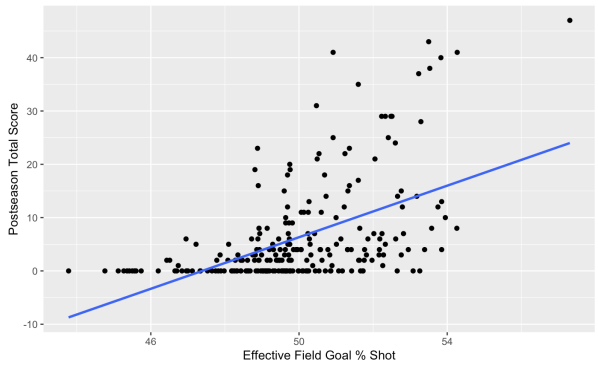
Field goal percentage is the ratio between the number of field goals attempted and the number of field goals made. A field goal is either a 2 point or 3 point shot taken during live play. This number is then

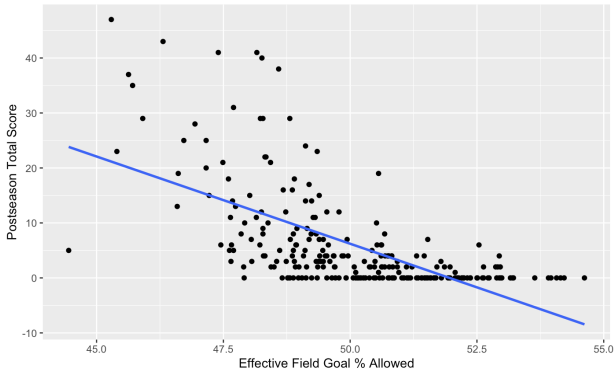
multiplied by 100 to give a percent known as field goal percentage. EFG_O represents the field goal percentage shot by a team. EFG_D represents the field goal percentage allowed by a team. Initially we fit two linear models to compare the relationship between field goal percentage and postseason score as well as field goal percentage allowed and postseason score. The relationships in the models were:

$$E[pst|\ EFG\_O] = \beta 0 + \beta 1 * EFG\_O$$

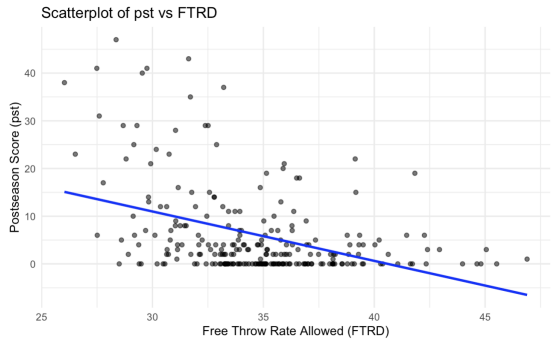
$$E[pst|\ EFG\_D] = \beta 0 + \beta 1 * EFG\_D$$

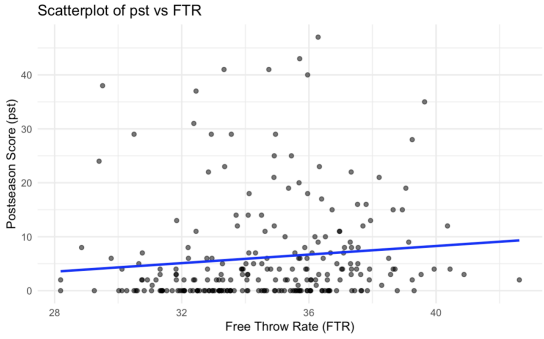
In the offensive regression model β0 is -114.54 which is not interpretable in this context as a team cannot have -114 post season points and no teams had a field goal percentage of 0. β1 in the offensive model is 2.417. This means that for every one percent increase in field goal percentage postseason score increases 2.42 points. In the defensive model β0 is 164.88 which is not interpretable in this context because this would mean a team who allowed a field goal percentage of 0 would have 164.88 postseason points, both of which do not make sense. β1 is -3.17 meaning that for every additional field goal percent allowed a teams total postseason points decreases by 3.17 points. All four of these coefficients are significant with a p-value of 2E-16. These models indicate that there is a relationship between both offense and postseason success and can be visualized through scatterplots.





Free Throws





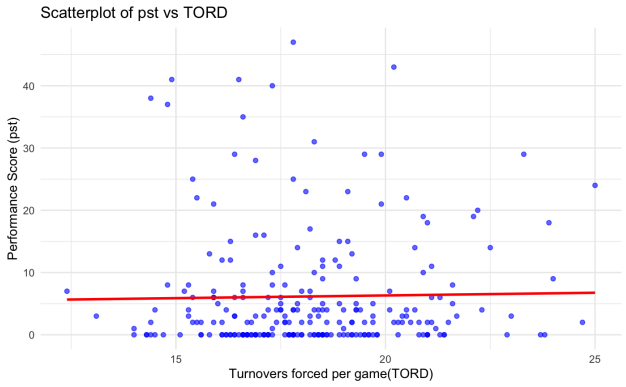
To explore the relationship between free throw rate and postseason score, we created two linear regression models. The first model assessed the impact of free throw rate allowed on postseason score and the second assessed free throw rate's impact on postseason score. The models gave us the equations:

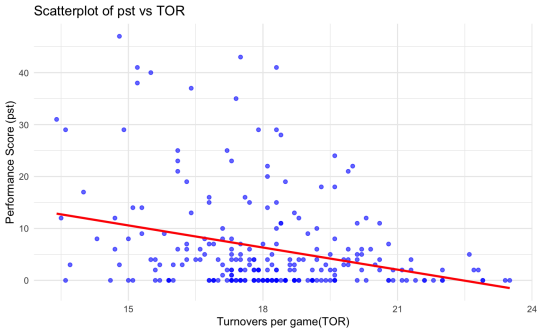
$$Postseason\ Score\ =\ 42.051\ -\ 1.0355\ *\ Free\ Throw\ Rate\ Allowed$$
$$Postseason\ Score\ =\ -\ 7.579\ +\ 0.396\ *\ Free\ Throw\ Rate$$

The regression for FTRD showed a strong negative association with postseason score, with a coefficient of -1.0355. This suggests that for every one-unit increase in FTRD, postseason score decreases by approximately 1.04 points, holding other factors constant. The regression for FTR showed a slightly positive association with postseason score, with a coefficient of 0.396. This suggests that for every one-unit increase in FTRD, postseason score decreases by approximately 0.396 points, holding other variables constant. The intercept of both models is rather insignificant because it represents the predicted postseason score when free throw rate or free throw rate allowed is 0. It is highly improbable for any team to either make or allow zero free throws over the course of a season.

Our analysis of free throw rate differential FTRD revealed a significant negative relationship with postseason success, suggesting that limiting opponent free throw opportunities—a defensive factor plays a more critical role than simply increasing offensive free throw attempts. This supports the idea that defensive performance, such as avoiding fouls and controlling the game's pace, may be more important than offensive dominance in achieving success in the NCAA tournament.

Turnovers





These two graphs show the Turnovers forced and committed per game plotted against postseason score. As you can see, with more turnovers committed, the post season score is lower, but with more turnovers forced, there is little change in postseason score. To figure out which variable had more of an impact on postseason success, I ran these 2 tests.

PostseasonScore=B0+B1*TORD

PostseasonScore=B0+B1*TOR

I found that the TORD coefficient is .087 while for TOR, the coefficient is -1.419.
This means that as the team committed one more turnover per game, we would expect their postseason score to decrease by 1.419 while we would expect, if we forced 1 more turnover per game, our postseason score would rise by .087. Unfortunately, only the TOR had a low p value, the TORD had a p value of .749, meaning that it is not statistically significant. This makes sense, because the coefficient of .087 is very low. The intercept is unimportant because there are gonna be no teams that commit 0 turnovers per game. Using these variables, this supports the idea that offense is more important than defense in the NCAA tournament.

Conference as a Confounder

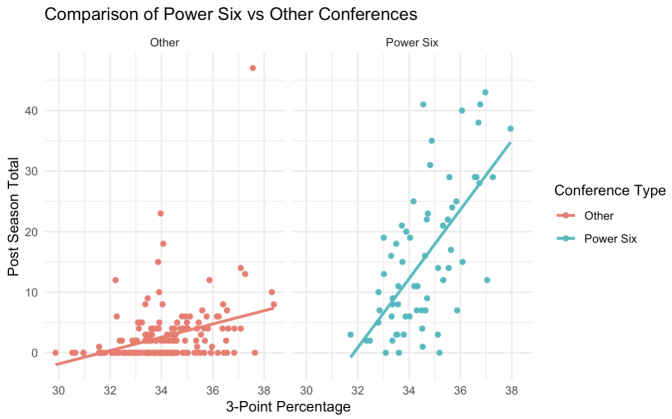
To further explore this relationship, we used conference as a confounding variable. The new linear regression equations are the following:

PostseasonScore = B0 + B1*OffensiveVariable + B2*ACCConference + B3*Big10Conference...

PostseasonScore = B0 + B1*DefensiveVariable + B2*ACCConference + B3*Big10Conference...

The ... at the end of the equations represent that there are numerous other conferences that have an impact on the relationship between the original variables in both equations. Through an initial analysis we found that many conferences don't have a relationship between variables and postseason score. However, the conferences that show a great effect between our variables and postseason score happen to be the power 6 conferences(ACC, Big10, Big12, SEC, Big East, Pac 12). The WCC also seems to show a significant relationship between the two variables, but the data is probably skewed because of Gonzaga's performance in the NCAA tournament. If you remove the outlier dot in each graph for the WCC, the trend line would probably be more flat. The impact of the power 6 conferences on the data will be explored in a further section.

So, to create the best visual for how conference confounds the relationship between each variable and postseason score, we compared the Power Six conferences to everyone else. An example graph of this is the following:



The other results for this can be found in the appendix, and all show that the power six conferences' stats have a greater impact on post season total.

Methods

To help us answer our research question, we created a few statistical models. The one that we used was the calculation of the z score for all the offensive predictors. We did this by first using the scale function on each of these predictors. Scale centers the column around the mean and standard deviation of the column. After scaling each of the predictors, we took a sum of all of these z scores for the original predictors. We then did the same process for the defensive predictors. After that, we had to create a linear regression model using both the z scores for defensive and offensive predictors. We also used conference as a confounder. To do this, we grouped 6 of the best conferences, the ACC, Pac 12, Big 10, Big 12, SEC, and the Big East. We did this because there is a difference in the postseason scores with these power six conferences, and the other conferences. The postseason scores for the power six conferences are significantly higher than the non power conferences. We listed a few of the other models that we considered beforehand in the data analysis, but we ultimately decided on this one because it took into account all of the offensive predictors and not just one. Our model looked something like this:

E[pst|Total_zscoreO ,Total_zscoreD ,PowerSix]=B0+B1*Total_scoreO+B2*Total_scoreD+B3*PowerSix

The confounder variable would be the Power six and the precision variables are the Total score for offense and defense. The null hypothesis is that there is no effect of offense or defense on the postseason score. This means that we would not see a difference between a good offensive team or a good defensive team and their post season success. Our alternative hypothesis is that there is a

relationship between offense or defense and postseason score. This means that we would expect a team with a good offense to do better or worse than a team with a good defense.

Results

The z score model produced the following results:

```
lm(formula = pst ~ Total_zscoreO * Total_zscoreD + PowerSix,
    data = cbbjoined)
Residuals:
     Min       1Q   Median       3Q      Max
-18.2822  -2.6818  -0.2336   1.6287  22.4427
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                   2.8092     0.4650   6.042 5.84e-09 ***
Total_zscoreO                 0.6735     0.1451   4.643 5.68e-06 ***
Total_zscoreD                -0.7019     0.1092  -6.429 6.95e-10 ***
PowerSixPower Six             8.9580     0.9022   9.929  < 2e-16 ***
Total_zscoreO:Total_zscoreD  -0.1658     0.0259  -6.401 8.12e-10 ***
```

Based on this model we can conclude that defense seems to have more of an effect on postseason success. The coefficient for Total_zscoreD is “higher” than that of Total_zscoreO which means that postseason score increases more for every additional defensive unit increase. Defense appears as a negative because when a team allows more on the defensive stats they perform worse. For example, if a team allowed a higher field goal percentage they would give up more points and may lose the game. A team that gives up a lower field goal percentage would therefore have less points scored against them and will have more success. This is why we can interpret the coefficients regardless of sign. Both coefficients are significant well below the p-value threshold of 0.05 and Total_zscoreD is even more significant than Total_zscoreO.

Conclusion

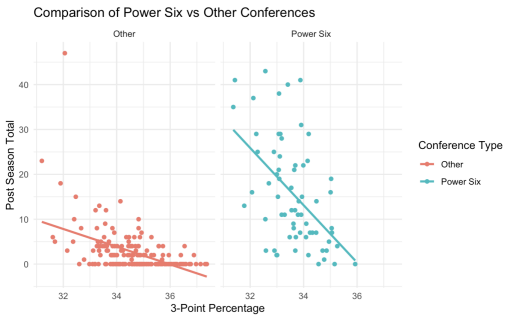
Based on our investigations, we can conclude that defense is more important than offense in March Madness. This would be important for the broader population because many people every March make brackets to try to make some money, unfortunately they usually fail at making any money. But

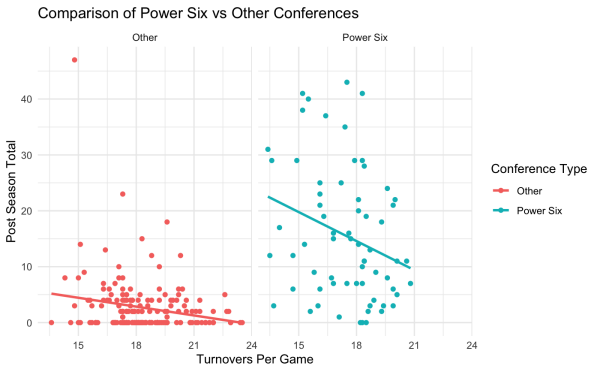
this research tells us that we should usually pick the team with the better defense rather than the better offense. There were not many limitations nor were there ethical concerns within this analysis. But there is a bias of using postseason scores, some may say that is not a true measure of success. We saw many teams with the same stats that didn't make the postseason. This is because those who select these teams are always going to pick the bigger conferences like SEC, ACC, and Big 10. We can see that there is a causal relationship between the defensive performance of a team and their post season success. An article from *The Medium* used similar data to what we had, but was able to come up with so many more graphs that explored a lot about college basketball. It shows that there is a lot you can do with college basketball data.

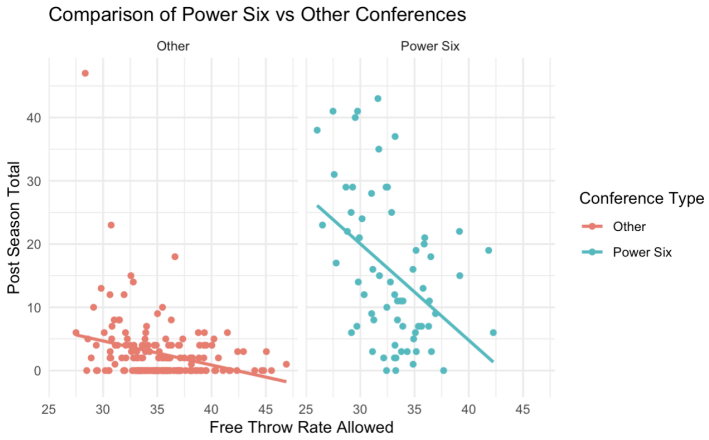
Appendix

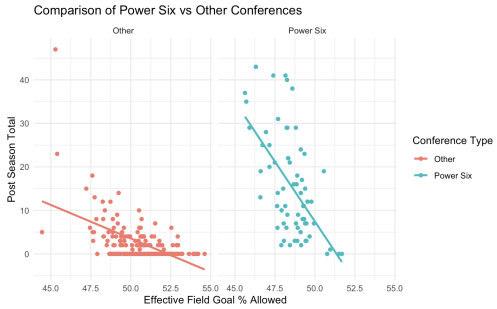
"College Basketball Dataset." n.d. Kaggle. Accessed December 8, 2024. https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset.

Chen, Kenneth. "College Basketball Data Analysis: Quantifying Winning with Box Score Statistics." *Medium*, Medium, 3 Mar. 2023, medium.com/@kennethylhs22/college-basketball-data-analysis-quantifying-winning-with-box-score-statistics-f16c68600d27.






Comparison of Power Six vs Other Conferences
Other
Power Six
Post Season Total
Free Throw Rate Allowed
Conference Type
Other
Power Six


Comparison of Power Six vs Other Conferences
Other
Power Six
Post Season Total
Effective Field Goal % Allowed
Conference Type
Other
Power Six