Analysis of March Madness

Olivia Wivestad, Conner Byrd, Dani Trejo, Aidan Gildea

due April 6, 2020

Introduction

As avid Duke Basketball fans and registered Cameron Crazies, all members of the Outliers team were disappointed to hear about the cancellation of March Madness. However, the unforeseen ending of Duke's season, along with numerous other colleges', is a necessary step in combating the ongoing Coronavirus pandemic. Across the globe, people have been forced to adjust their habits and put their lives on hold. With shelter in place and lockdown orders being increasingly mandated, we have found ourselves with more free time than ever before. With this, we are oftentimes left wondering "what if" or "what would have happened." In most cases, these questions are left unanswered; however, the outcome of this year's March Madness may be one we can answer.

Drawing from data of Division I basketball teams in past seasons and using R, our team plans to answer the primary research question: **What statistical factors most attribute to success in March Madness for teams?** Our resulting findings will hopefully uncover the true influence of "madness" on the tournament and offer a prediction of who the 2020 championship team may have been.

Our dataset was provided on Kaggle. The Kaggle user scraped the data from another basketball statistic dataset found here. The site is run by Bart Torvik and colleagues, and it records basketball statistics from the past 6 years. Most of the figures are drawn from various sites, including the official NCAA site, after each official game and season. Websites like the official NCAA website provide detailed statistics on teams, games, and seasons, as well as some figures for an extra fee. The Kaggle user cleaned Torvik's dataset and added three additional variables (POSTSEASON, SEED, and YEAR), giving us our dataset. Each case represents a team from a specific season (2015-2020) and consists of the following variables:

Variables	Descriptions
Team	The Division I college basketball school
CONF	The Athletic Conference in which the school participates in
G	Number of games played
W	Number of games won
ADJOE (Adjusted Offensive Efficiency)	An estimate of the offensive efficiency (points scored per 100 possessions) a team would have against the average Division I defense
ADJDE (Adjusted Defensive Efficiency)	An estimate of the defensive efficiency (points allowed per 100 possessions) a team would have against the average Division I offense)
BARTHAG	Power Rating (Chance of beating an average Division I team)
EFG_O	Effective Field Goal Percentage Shot
EFG_D	Effective Field Goal Percentage Allowed
TOR	Turnover Rate
TORD	Steal Rate
ORB	Offensive Rebound Percentage
DRB	Defensive Rebound Percentage
FTR	Free Throw Rate (How often the given team shoots Free Throws)
FTRD	Two-Point Shooting Percentage
2P_O	Two-Point Shooting Percentage Allowed
3P_O	Three-Point Shooting Percentage
ADJ_T	Three-Point Shooting Percentage Allowed
WAB (Wins Above Bubble)	The bubble refers to the cut off between making the NCAA March Madness Tournament and not making it
POSTSEASON	Round where the given team was eliminated or where their season ended + R68 = First Four, R64 = Round of 64, R32 = Round of 32, S16 = Sweet Sixteen, E8 = Elite Eight, F4 = Final Four, 2ND = Runner-up, Champion = Winner of the NCAA March Madness
SEED	Seed in the NCAA March Madness Tournament
YEAR	Season

Data Analysis Plan

To answer our main research question, we will use the following variables:

Dependent: POSTSEASON Independent: We will use all of the other variables to analyze our dependent variable, however, we expect many of them to be unimportant. Using backwards elimination, we will remove the extraneous variables. We initially predict that the variables SEED, WAB, G, W, ADJOE, ADJDE, and BARTHAG will be most helpful in answering our question.

To start, much of our statistical inferencing will be based on creating a linear model with backward selection to find equation of adequate predictors. Doing this would eliminate unimportant variables from our dataset and give us a more focused idea on what to model. We will observe which variables have the strongest influence on a team's chance of making the Final Four in the tournament by looking at their respective p-values. Basic descriptive methods can be used to summarize these findings. We could also use this to input the statistics of teams from the 2019-20 basketball season and predict which teams would make the Final Four and beyond, hopefully giving us some closure on this sad ending of a basketball season.

Who has won in the past?

TEAM	POSTSEASON	YEAR	SEED	CONF
Virginia	Champions	2019	1	ACC
Villanova	Champions	2018	1	BE
North Carolina	Champions	2017	1	ACC
Villanova	Champions	2016	2	BE
Duke	Champions	2015	1	ACC

There are four teams who won the NCAA in the last five years: Villanova won in both 2016 and 2018. Since these teams won, we can use statistics from their respective seasons to discern what variables have the biggest effect on a winning run in the NCAA tournament.

What conferences produce the most successful teams?

CONF	n
ACC	5
B10	4
B12	3
SEC	3
BE	2
MVC	1
P12	1
WCC	1

There are only eight conferences with teams that have made it to the final four in the past 5 years out of a total 32 conferences. The ACC leads all conferences, having 5 total appearances. This reality could play an important part in narrowing down what team would have won in 2020.

New Variables and Datasets

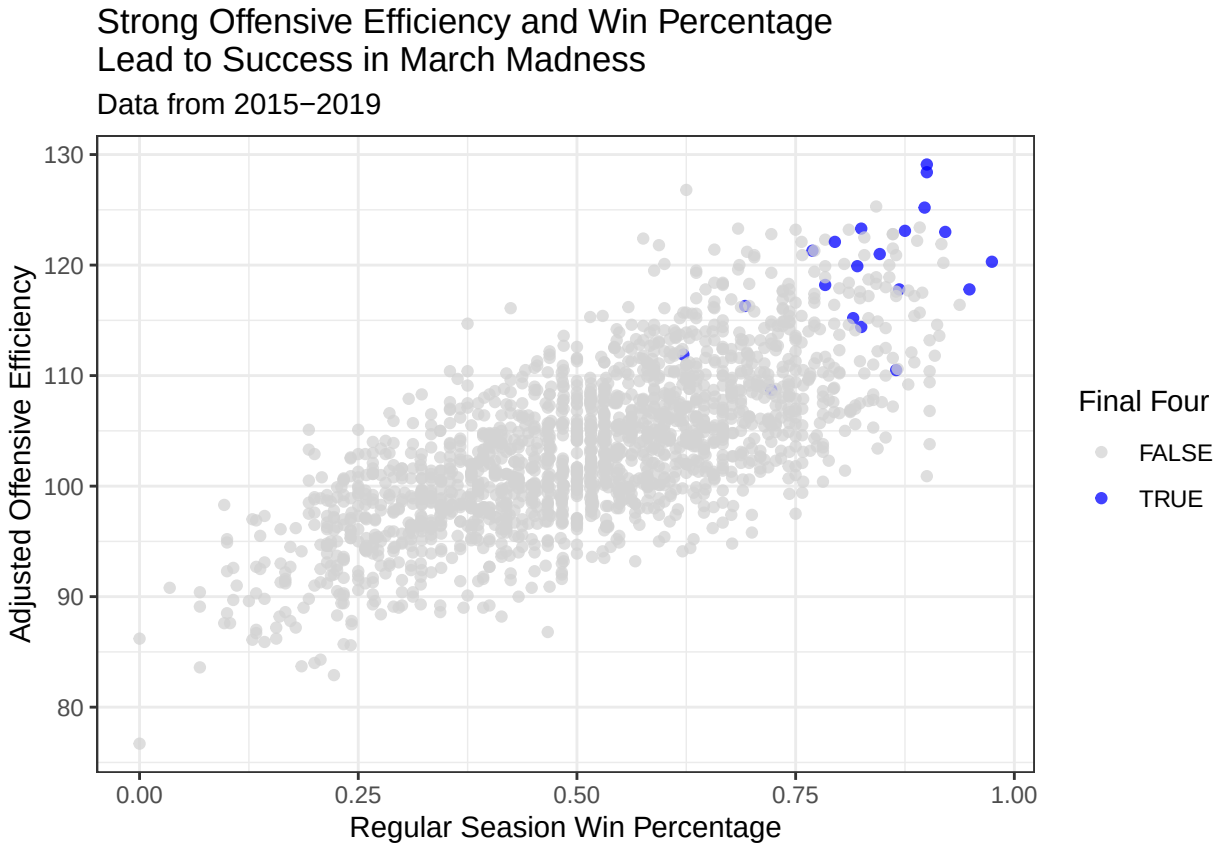
We added two variables: "final_four"and "post_wp" which denotes whether a team made it to the Final Four or better in a specific season and the win percentage variable represent the proportion of games a team won in the regular season, respectfully.

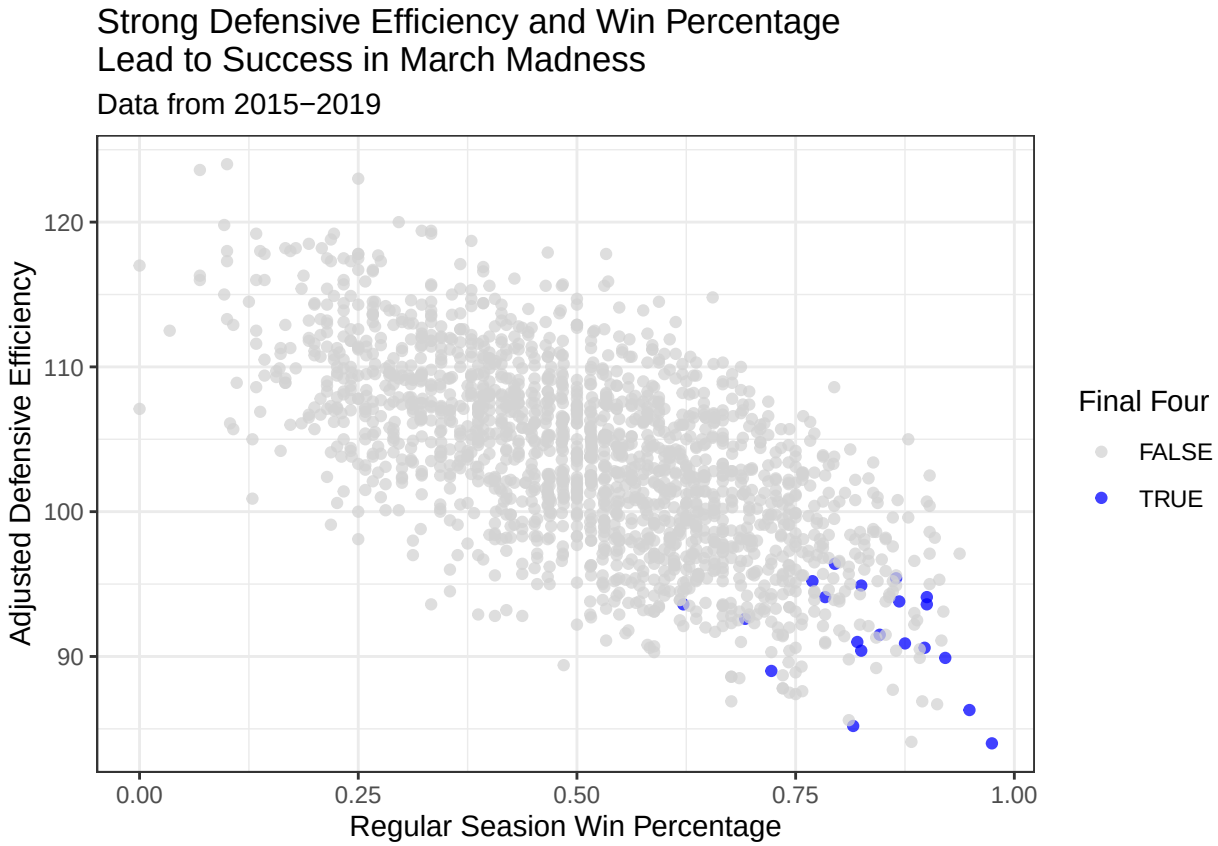
Duke's Winning Season

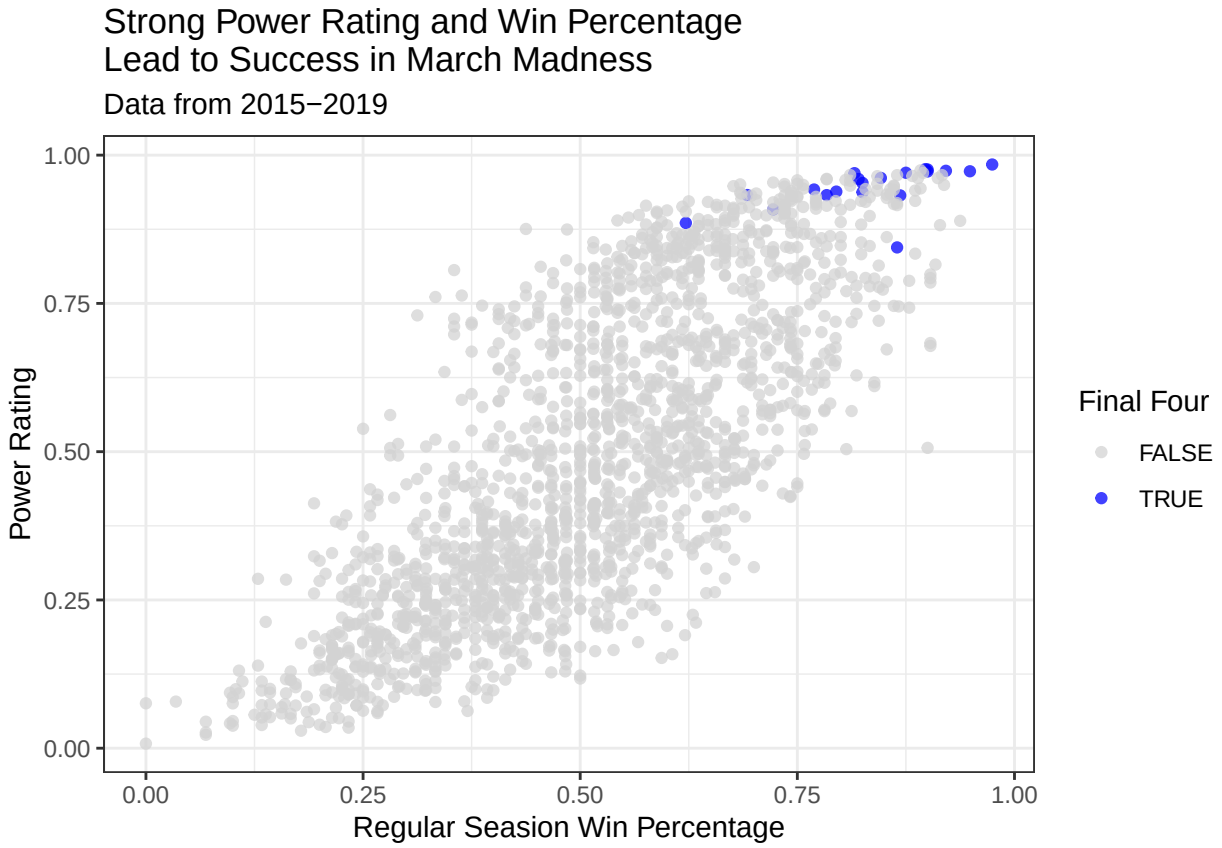
SEED	WAB	win_percent	ADJOE	ADJDE	BARTHAG
1	10.7	0.8974359	125.2	90.6	0.9764

Throughout the season, the Cameron Crazies have hoped for a sixth NCAA Championship. We will never know how this year's March Madness would have gone for Duke; however, looking at statistics from 2015 championship team may show us what factors led to their success.

Visualizations


Strong Offensive Efficiency and Win Percentage
Lead to Success in March Madness
Data from 2015−2019
Adjusted Offensive Efficiency
130
120
110
100
90
80
0.00
0.25
0.50
0.75
1.00
Regular Season Win Percentage
Final Four
FALSE
TRUE


Strong Defensive Efficiency and Win Percentage
Lead to Success in March Madness
Data from 2015−2019
Adjusted Defensive Efficiency
120
110
100
90
0.00
0.25
0.50
0.75
1.00
Regular Season Win Percentage
Final Four
FALSE
TRUE



Glimpse

```
## Observations: 1,757
## Variables: 27
## $ TEAM        <chr> "North Carolina", "Wisconsin", "Michigan", "Texas Tech"...
## $ CONF        <chr> "ACC", "B10", "B10", "B12", "WCC", "ACC", "ACC", "ACC",...
## $ G           <dbl> 40, 40, 40, 38, 39, 39, 38, 39, 40, 40, 36, 38, 36, 37,...
## $ W           <dbl> 33, 36, 33, 31, 37, 35, 35, 33, 35, 36, 27, 32, 24, 29,...
## $ ADJOE       <dbl> 123.3, 129.1, 114.4, 115.2, 117.8, 125.2, 123.0, 121.0,...
## $ ADJDE       <dbl> 94.9, 93.6, 90.4, 85.2, 86.3, 90.6, 89.9, 91.5, 90.9, 9...
## $ BARTHAG     <dbl> 0.9531, 0.9758, 0.9375, 0.9696, 0.9728, 0.9764, 0.9736,...
## $ EFG_O       <dbl> 52.6, 54.8, 53.9, 53.5, 56.6, 56.6, 55.2, 51.7, 56.1, 5...
## $ EFG_D       <dbl> 48.1, 47.7, 47.7, 43.0, 41.1, 46.5, 44.7, 48.1, 46.7, 4...
## $ TOR         <dbl> 15.4, 12.4, 14.0, 17.7, 16.2, 16.3, 14.7, 16.2, 16.3, 1...
## $ TORD        <dbl> 18.2, 15.8, 19.5, 22.8, 17.1, 18.6, 17.5, 18.6, 20.6, 1...
## $ ORB         <dbl> 40.7, 32.1, 25.5, 27.4, 30.0, 35.8, 30.4, 41.3, 28.2, 2...
## $ DRB         <dbl> 30.0, 23.7, 24.9, 28.7, 26.2, 30.2, 25.4, 25.0, 29.4, 2...
## $ FTR         <dbl> 32.3, 36.2, 30.7, 32.9, 39.0, 39.8, 29.1, 34.3, 34.1, 2...
## $ FTRD        <dbl> 30.4, 22.4, 30.0, 36.6, 26.9, 23.9, 26.3, 31.6, 30.0, 2...
## $ `2P_O`      <dbl> 53.9, 54.8, 54.7, 52.8, 56.3, 55.9, 52.5, 51.0, 57.4, 5...
## $ `2P_D`      <dbl> 44.6, 44.7, 46.8, 41.9, 40.0, 46.3, 45.7, 46.3, 44.1, 4...
## $ `3P_O`      <dbl> 32.7, 36.5, 35.2, 36.5, 38.2, 38.7, 39.5, 35.5, 36.2, 4...
## $ `3P_D`      <dbl> 36.2, 37.5, 33.2, 29.7, 29.0, 31.4, 28.9, 33.9, 33.9, 3...
## $ ADJ_T       <dbl> 71.7, 59.3, 65.9, 67.5, 71.5, 66.4, 60.7, 72.8, 66.7, 6...
## $ WAB         <dbl> 8.6, 11.3, 6.9, 7.0, 7.7, 10.7, 11.1, 8.4, 8.9, 10.6, 5...
## $ POSTSEASON  <chr> "2ND", "2ND", "2ND", "2ND", "2ND", "Champions", "Champi...
```

```
## $ SEED        <dbl> 1, 1, 3, 3, 1, 1, 1, 1, 2, 1, 4, 3, 6, 1, 2, 9, 1, 3, 1...
## $ YEAR        <dbl> 2016, 2015, 2018, 2019, 2017, 2015, 2019, 2017, 2016, 2...
## $ win_percent <dbl> 0.8250000, 0.9000000, 0.8250000, 0.8157895, 0.9487179, ...
## $ final_four  <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, T...
## $ post_wp     <dbl> 0.833, 0.833, 0.833, 0.833, 0.833, 1.000, 1.000, 1.000,...
```

```
## Observations: 353
## Variables: 22
## $ RK      <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ...
## $ TEAM    <chr> "Kansas", "Baylor", "Gonzaga", "Dayton", "Michigan St.", "D...
## $ CONF    <chr> "B12", "B12", "WCC", "A10", "B10", "ACC", "BE", "B10", "ACC...
## $ G       <dbl> 30, 30, 33, 31, 31, 31, 30, 31, 31, 31, 31, 31, 32, 31, 31,...
## $ W       <dbl> 28, 26, 31, 29, 22, 25, 24, 21, 24, 30, 24, 23, 21, 19, 21,...
## $ ADJOE   <dbl> 116.1, 114.5, 121.3, 119.5, 114.8, 115.3, 120.6, 114.6, 115...
## $ ADJDE   <dbl> 87.7, 88.4, 94.3, 93.4, 91.3, 91.9, 96.4, 92.6, 93.9, 92.8,...
## $ BARTHAG <dbl> 0.9616, 0.9513, 0.9472, 0.9445, 0.9326, 0.9310, 0.9289, 0.9...
## $ EFG_O   <dbl> 53.7, 49.4, 57.5, 59.7, 52.6, 52.6, 55.2, 52.3, 52.5, 54.6,...
## $ EFG_D   <dbl> 43.7, 45.2, 47.6, 46.6, 43.3, 45.7, 48.4, 46.2, 45.1, 45.2,...
## $ TOR     <dbl> 18.7, 17.8, 15.3, 18.0, 18.1, 17.8, 15.9, 19.1, 18.0, 16.2,...
## $ TORD    <dbl> 18.6, 22.7, 18.4, 18.8, 15.8, 20.2, 17.6, 18.3, 17.4, 21.3,...
## $ ORB     <dbl> 32.6, 35.8, 33.6, 26.4, 32.8, 34.8, 23.9, 31.1, 32.0, 28.2,...
## $ DRB     <dbl> 26.4, 29.8, 22.7, 26.6, 26.0, 28.0, 30.2, 25.5, 25.0, 25.4,...
## $ FTR     <dbl> 35.8, 30.8, 38.8, 33.9, 30.8, 35.6, 28.8, 36.7, 32.2, 28.3,...
## $ FTRD    <dbl> 23.2, 30.8, 21.8, 30.9, 29.3, 30.9, 23.4, 29.3, 29.0, 30.6,...
## $ `2P_O`  <dbl> 54.9, 47.5, 57.4, 62.3, 52.9, 52.5, 53.0, 49.7, 50.1, 53.0,...
## $ `2P_D`  <dbl> 42.4, 44.4, 47.4, 45.1, 43.4, 46.0, 48.9, 44.2, 45.1, 45.3,...
## $ `3P_O`  <dbl> 34.1, 35.1, 38.6, 37.1, 34.8, 35.2, 38.7, 37.3, 37.6, 37.9,...
## $ `3P_D`  <dbl> 30.5, 31.1, 32.0, 33.0, 28.7, 29.9, 31.8, 32.7, 30.1, 30.1,...
## $ ADJ_T   <dbl> 67.4, 66.2, 72.0, 67.5, 69.3, 71.7, 68.3, 66.2, 66.9, 64.7,...
## $ WAB     <dbl> 10.8, 8.5, 7.7, 6.8, 5.2, 5.1, 6.1, 3.8, 4.3, 6.7, 6.8, 2.7...
```

Methods and Results

Clarifications

For the goal of our analysis, we define success in the March Madness tournament by a team's ability to make the Final Four round, in turn, meaning they have a high Predicted Post Season Win Percentage.

In order to make predictions about what would have happened in the 2020 tournament, we must use data from the 2015-2019 seasons and their respective March Madness tournaments to create appropriate prediction models.

Identifying which variables are the greatest predictors.

Term	Estimate	P-value
(Intercept)	1.4017446	0.0060349
ADJOE	0.0425383	0.0000000
ADJDE	-0.0491725	0.0000000
BARTHAG	-1.3773011	0.0000030

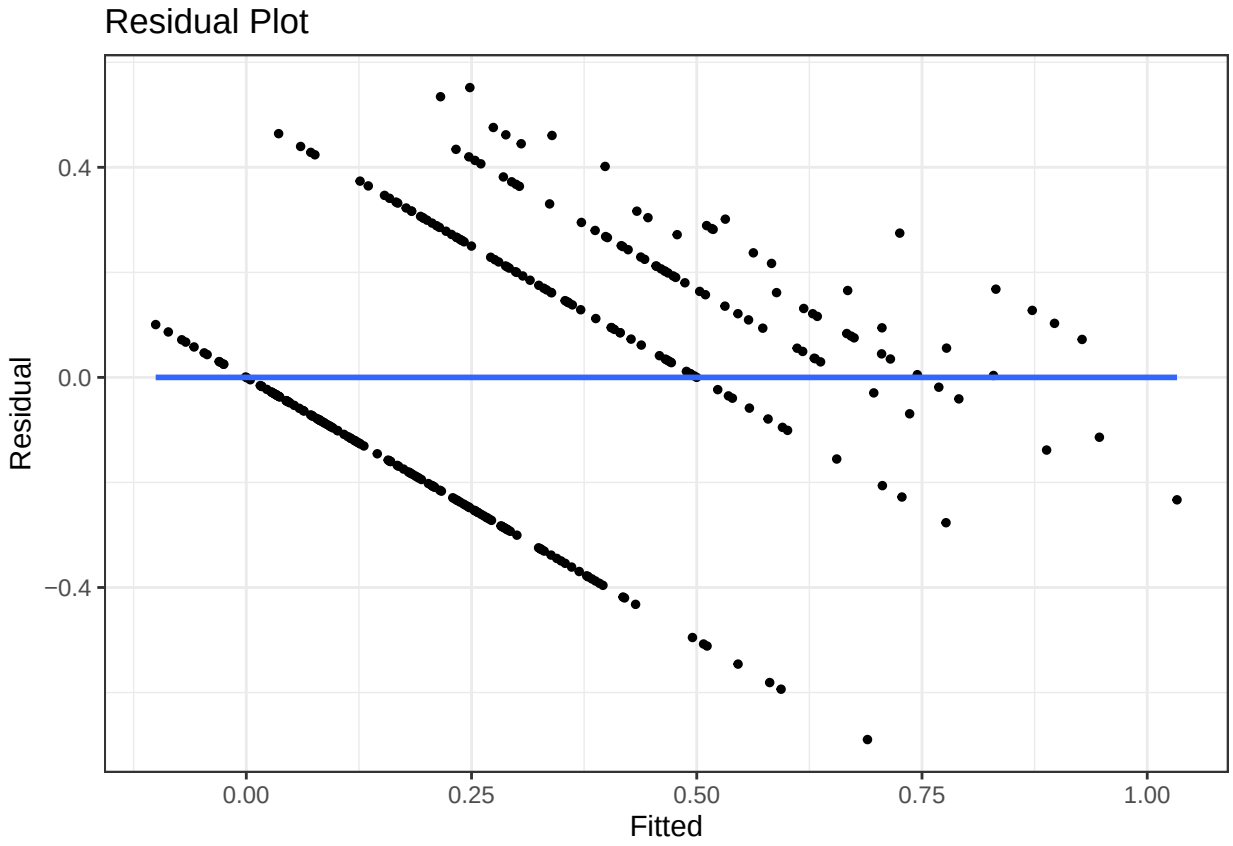
R-squared
0.4687232

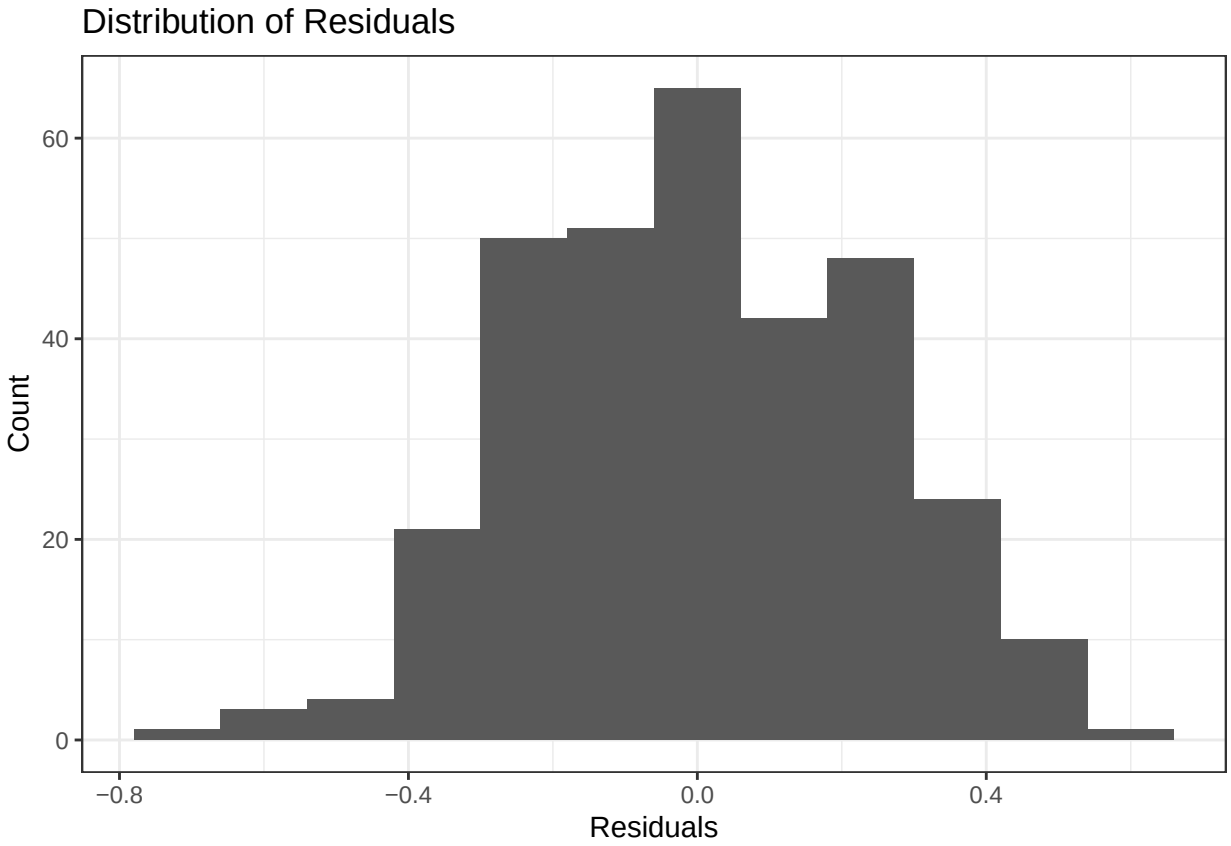
The equation for the linear model above is:

Predicted Post Season Win % = 1.40+0.0425(ADJOE)-0.0492(ADJDE)-1.378(BARTHAG)

With a r-squared value of 0.4687, we can attribute 46.87% of the variability in the Predicted Post Season Win Percentage to the three predictors: ADJOE, ADJDE, and BARTHAG. This is relatively low, but as we will explain in the Discussion section, this could be a sign of the randomness March Madness is known for.

Conditions





The linearity and variance assumptions of the linear model are not satisfied by the residual plot because it does not display a random pattern. The residual plot seems to have different groupings of datapoints, the result of the post-season win percentages being ordinal. The distribution of the residuals is approximately normal, allowing us to satisfy the normality assumption for the linear model. We assume that independence is satisfied because each team is unique – whether it be a different season or different college. While the violation of the linearity and variance assumption would typically nullify the model, we are using it for variable selection, and not inference, therefore, justifying its use. It is important to mention this model will not be used for any hypothesis testing or its p-values, instead it is only a means for supporting the ad hoc process of variable selection.

Reasoning for Linear Model and Interpretation

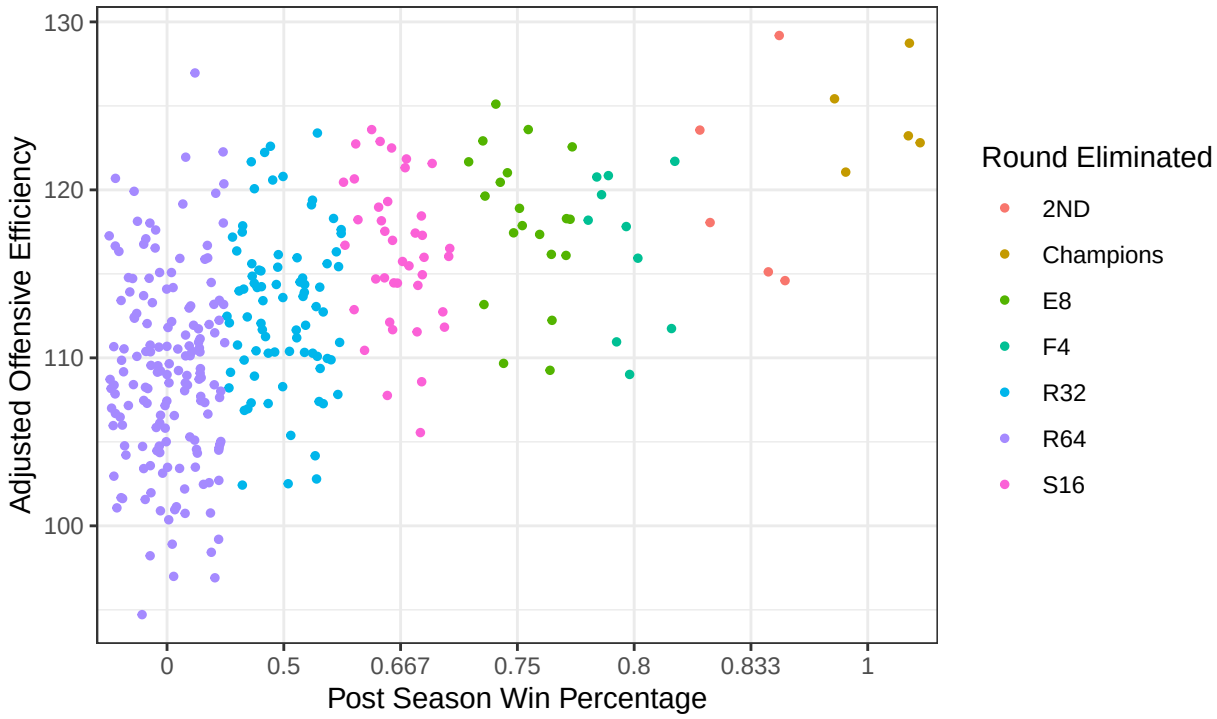
By using backward selection, we were able to remove the least contributive predictors, leaving us only with those which we consider to be the most accurate for predicting Post Season Win Percentage. This left us with three variables:
1. ADJOE: For each unit increase in a team's Adjusted Offensive Efficiency, we expect their predicted Post Season Win Percentage to increase by .0425 percentage points.
2. ADJDE: For each unit increase in a team's Adjusted Defensive Efficiency, we expect their predicted Post Season Win Percentage to decrease by .0492 percentage points.
3. BARTHAG: For each unit increase in a team's Power Rating, we expect their predicted Post Season Win Percentage to decrease by 1.378 percentage points.

Predictor Visualizations

To ensure that our selected predictors exhibit the desired relationship with Post-Season Win Percentage, we created visualization below. All data points are categorized by the round they were eliminated during in March Madness, giving a better picture of how each predictor influences postseason prospects.

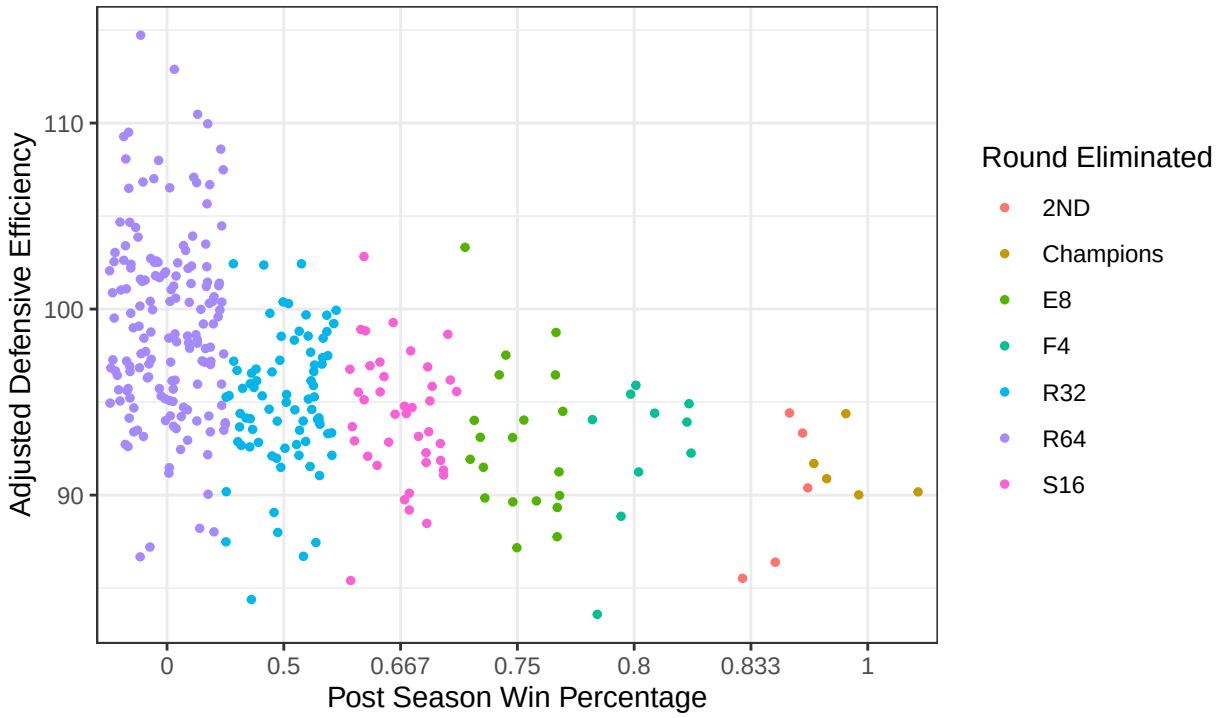
Offensive Efficiency Positively Correlates with Post Season Win Percentage

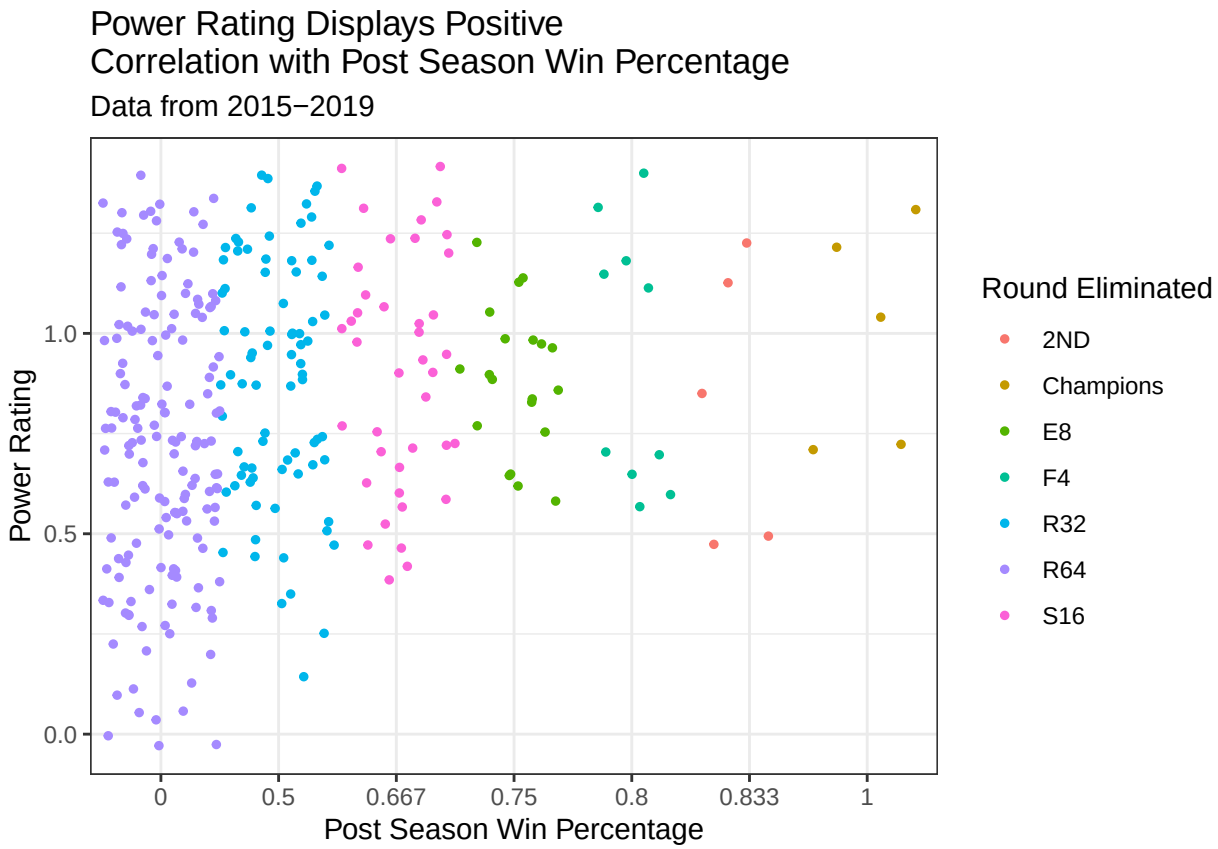
Data from 2015−2019



Defensive Efficiency Negatively Correlates with Post Season Win Percentage

Data from 2015−2019





All the visualizations display the desired relationship between each predictor and Post Season Win Percentage. For both ADJOE and BARTHAG, this is a positive relationship, and for ADJDE, it is a negative relationship.

Creating a logistic regression.

Now that we have identified the best predictors for Predicted Post Season Win Percentage, we can use these variables to attempt to classify which teams (Name+Year) made it to the Final Four in a March Madness tournament. We can use a Logistic Regression model to assess the test accuracy using the three variables which we have identified as the best predictors for Predicted Post Season Win Percentage (ADJOE, ADJDE, and BARTHAG).

Prediction Accuracy
0.94

F1 Score
0.9684211

The logistic regression model predicts if a team made it to the Final Four with an accuracy of 94%. However, this high prediction accuracy is likely a result of the large number of teams that do not make the Final Four (60/64), which already accounts for 93.75% of the accuracy. The F1 Score of the model, 0.968, is a better measure because it takes into account all four outcomes of a model:

- false-positive (predicted to make Final Four when in reality they did not)
- false-negative (predicted to not make Final Four when in reality they did)
- true-positive (predicted to make Final Four and they did in reality)
- true-negative (predicted to not make Final Four when in reality they did not)

This high F1 score (near 1) confirms that these three variables (ADJOE, ADJDE, and BARTHAG) accurately

predict whether a team is likely to make it to the Final Four round of March Madness. By "tuning" the model, we determined that a probability threshold of 40% produced the most accurate F1 score and was best for classifying the teams.

Creating a KNN-model.

We can also use a KNN-model to classify whether a team is predicted to make it to the Final Four round. As concluded in the Logistic Regression model, the F1 score is a better measure of test accuracy; therefore, we will use the F1 Score to assess which k-nearest neighbor number results in the best model.

F1 Score	k
0.9583333	5

The KNN-model has the highest test accuracy when using a k-nearest neighbor number of 5, which produces a KNN model with a F1 score of 0.958. This high F1 score (near 1) confirms that these three variables (ADJOE, ADJDE, and BARTHAG) accurately help predict whether a team is likely to make it to the Final Four round of March Madness.

Comparison of KNN and Logit Models

After generating both a Logistic Regression model and a KNN-model with a k-nearest neighbor of 5, we observe that the logistic model has a higher F1 score, and therefore, is the more accurate model. We can conclude that both types of models – when using explanatory variables ADJOE, ADJDE, BARTHAG – are highly acccurate at classifying whether or not a team made it to the Final Four.

Would Duke have made it to the Final Four in 2020?

The ADJOE, ADJDE, and BARTHAG of this year's Duke basketball team will predict whether they would have made it to the Final Four. This dataset was sourced from the same website and user as the other seasons, and provides all the same statistics for 2020 teams.

With the logistic regression model having a higher prediction accuracy with an F1 score of 0.968, we will use it over the KNN-model for this analysis.

Predicted Probability
0.071

Final Four?
FALSE

According to our logistic regression model, the 2019-2020 Duke Men's Basketball team would not have made it to the Final Four.

But March Madness is known for is wild outcomes, so let's see how our model has distinctly predicted Duke's success in the 2015-2019 seasons.

Have predictions stopped Duke Basketball from dominating over the past five years?

Since we refused to believe that Duke would have been eliminated before the Final Four this year, we decided to analyze how accurate our model has been in predicting Duke Final Four appearances over the last five years.

Year	Predicted Probability	Predicted Final Four	Prediction: TRUE or FALSE?
2015	0.719	TRUE	TRUE
2018	0.332	FALSE	TRUE
2019	0.381	FALSE	TRUE
2017	0.164	FALSE	TRUE
2016	0.052	FALSE	TRUE

Prediction Accuracy
1

The logistic regression model 100% accurately predicted whether or not Duke made it to the Final Four in the 2015-2019 seasons, which provides confidence in the conclusion that Duke Men's Basketball team would not have made it to the Final Four in 2020. Unfortunately, our analysis has predicted that we would not have seen the results we all hoped for, which was a Final Four appearance for the Blue Devils.

What teams were most likely to make the Final Four in 2020?

Team	Predicted Probability
Kansas	0.296
Gonzaga	0.174
Baylor	0.166
Dayton	0.142
Creighton	0.076
Michigan St.	0.075
Duke	0.071
Ohio St.	0.047
Louisville	0.035
San Diego St.	0.035

By using our logistic regression model and using data from the 2020 season, we were able to calculate the probability that each team would make the Final Four in a hypothetical 2020 March Madness. While we have the top 10 most likely Final Four teams above, it looks like the Final Four this year would have included: Kansas, Gonzaga, Baylor, and Dayton. Since Kansas has the highest probability of making to the Final Four, we have evidence to believe that they had the best odds of winning the tournament.

Discussion

With a general knowledge of basketball and the initial visualizations in our introduction, our group was able to narrow down the variables of interest. However, to further ensure confidence in our predictors, and in turn, our models, we created a linear model with backward selection. This resulted in three valuable predictors for Predicted Post Season Win Percentage: ADJOE, ADJDE, and BARTHAG. When used as explanatory variables in both a KNN-model and Logistic Regression model, postseason teams were classified by whether or not they made the final four. By comparing the models' F1 scores, we concluded that the Logistic Regression model was the most accurate in predicting which teams made the Final Four in the 2015-2019 seasons. Since its F1 score was near perfect at 0.979, we have deduced that the three predictors – ADJOE, ADJDE, and BARTHAG – largely attribute to success in March Madness.

Using this information from the 2015-2019 seasons, we were able to predict the postseason prospects of the 2020 Duke team. The Logistic Regression model predicted an unsuccessful outcome for the Brotherhood, meaning, they would not have made the Final Four. To further investigate this conclusion, we completed another Logistic Regression model, classifying whether only the past Duke teams had made the Final Four. The purpose of training the model without the Duke teams was to see if they were an outlier in our original

model. If they were an anomaly within the original, than this test would produce a low prediction accuracy. However, the model had a prediction accuracy of 100%, therefore substantiating the conclusion that our model was correct and Duke would not have made the Final Four this year.

We also used data from the 2020 season to predict which teams were most likely to have made the Final Four if March Madness had not been cancelled. Our Logistic Regression model predicted that (in descending order) Kansas, Gonzaga, Baylor, and Dayton were most likely to comprise the Final Four. Since Kansas had the highest predicted probability to make the Final Four, at 0.296, we believe that they would have been the winner of March Madness 2020 (sad!).

As explained before, we initially narrowed the variables to those we thought to be potentially meaningful with the hope of creating a functioning linear model with backward selection. Had we kept these variables, we might have ended up with a larger number of statistically significant predictors, deeming our current analysis not entirely comprehensive. Additionally, the linear model's relatively low r-squared value of 0.4687 is not the level of fit we would have hoped for, but it may be a sign that March Madness is deserving of its name. The high prediction accuracy of the Logistic Regression model was most likely a result of the limited number of true Final Four teams, thus why we measured the test accuracy using the F1 score instead.

Looking back at our prediction of whether or not Duke would have been in the Final Four, it makes sense that the Logistic Regression model classified that they would not make it because no team in our 2020 Logistic Regression model reached the probability threshold of 0.4. In reality, Duke is the 7th most likely team to reach the Final Four based on our full 2020 model, giving the Cameron Crazies some satisfying closure to a shortened season.

Reference

https://www.rdocumentation.org/packages/MLmetrics/versions/1.1.1/topics/F1_Score

http://haozhu233.github.io/kableExtra/awesome_table_in_pdf.pdf