March Madness and Gambling

Mac McLean, Santhosh Rajendran & Ariana Arenson

BACKGROUND

- March Madness Tournament is an exciting time for athletes, fans, and sports betters
- Approximately $9 billion wagered during 2021 March Madness
- Over/Under betting is popular form of betting - predicting the total number of combined points based on a line

Our Mission: To see if there is a way to gain an edge in predicting the total number of points in a game through multiple linear regression of team's statistics

Research Questions:

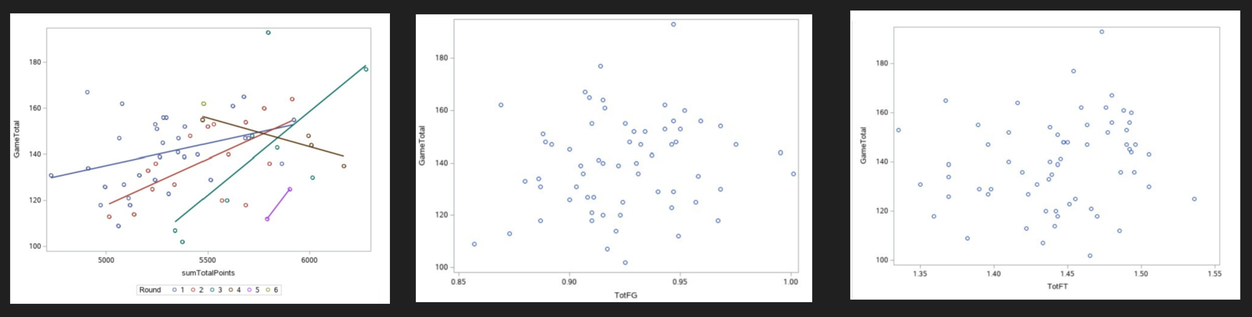
1. Do higher season long-term statistics lead to a larger total number of points in a March Madness game?
2. Do shooting specific statistics (such as 3P%, FT%, etc.) correlate with the total amount of points in a March Madness game?
3. Do non-team statistics, such as seed, round, or region correlate with the total amount of points in a March Madness game?

Our Data:

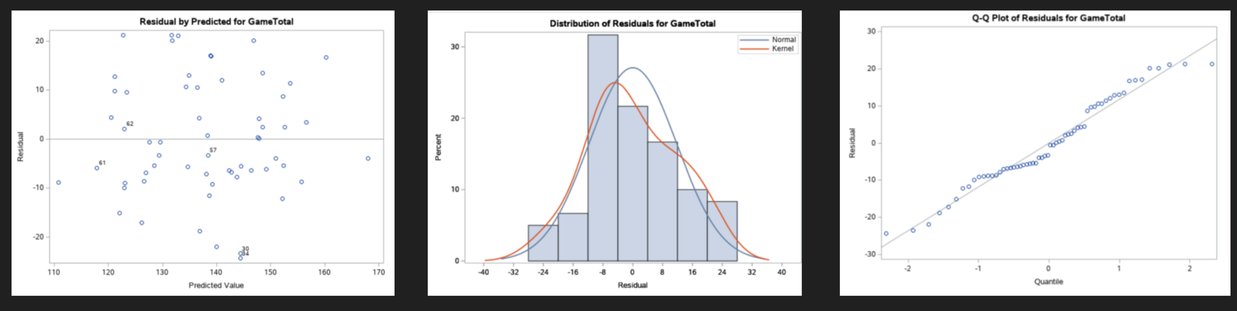
- collected from sportsreference.com, which is the most reliable source for all kinds of sports statistics data
- combined the team stats with a historical dataset from dataworld.com that has the results of all of the games from the 2019 tournament
- focus on the sum of team stats for each game

Variables: Round, Region Name, Seed1, Seed2, GameTotal, sumTotalPoints, sumPointsAllowed, fgPer, threeptPer, ftPer, sumRebounds, sumAssists

EXPLORATORY DATA ANALYSIS



REGRESSION ASSUMPTIONS



STEP 1: Add Quantitative Predictors

Initial: $\text{GameTotal} = \beta_0 + \beta_1\text{fgPer} + \beta_2\text{ftPer} + \beta_3\text{sumTotalPoints} + \beta_4\text{threePtPer} + \beta_5\text{sumAssists} + \beta_6\text{sumRebounds} + \beta_7\text{sumPointsAllowed} + \beta_8\text{sumSeed}$

Final: $\text{GameTotal} = \beta_0 + \beta_1\text{threePtPer} + \beta_2\text{sumRebounds} + \beta_3\text{sumPointsAllowed}$

FINAL MODEL FOR STEP 1

Adj. R-Squared: .4361
P-value: <.0001
Root MSE: 12.655

STEP 2: Add Qualitative Predictors

Initial: $\text{GameTotal} = \beta_0 + \beta_1\text{threePtPer} + \beta_2\text{sumRebounds} + \beta_3\text{sumPointsAllowed} + \beta_4\text{Round} + \beta_5\text{East} + \beta_6\text{South} + \beta_7\text{West} + \beta_8\text{Midwest} + \beta_9\text{Seed1} + \beta_{10}\text{Seed2}$

Final: $\text{GameTotal} = \beta_0 + \beta_1\text{threePtPer} + \beta_2\text{sumRebounds} + \beta_3\text{sumPointsAllowed} + \beta_4\text{Round} + \beta_5\text{Seed2}$

FINAL MODEL FOR STEP 2

Adj. R-squared: .4697
P-value: <.0001
Root MSE: 12.157

STEP 3: Add Interactions

Initial: $\text{GameTotal} = \beta_0 + \beta_1\text{threePtPer} + \beta_2\text{sumRebounds} + \beta_3\text{sumPointsAllowed} + \beta_4\text{Round} + \beta_5\text{threePtPer}*\text{Seed2} + \beta_6\text{threePtPer}*\text{Round} + \beta_7\text{threePtPer}*\text{sumPointsAllowed} + \beta_8\text{threePtPer}*\text{sumRebounds} + \beta_9\text{sumRebounds}*\text{Seed2} + \beta_{10}\text{sumRebounds}*\text{Round} + \beta_{11}\text{sumRebounds}*\text{sumPointsAllowed} + \beta_{12}\text{sumPointsAllowed}*\text{Seed2} + \beta_{13}\text{sumPointsAllowed}*\text{Round} + \beta_{14}\text{Round}*\text{Seed}$

No Interactions Were Found to be Significant

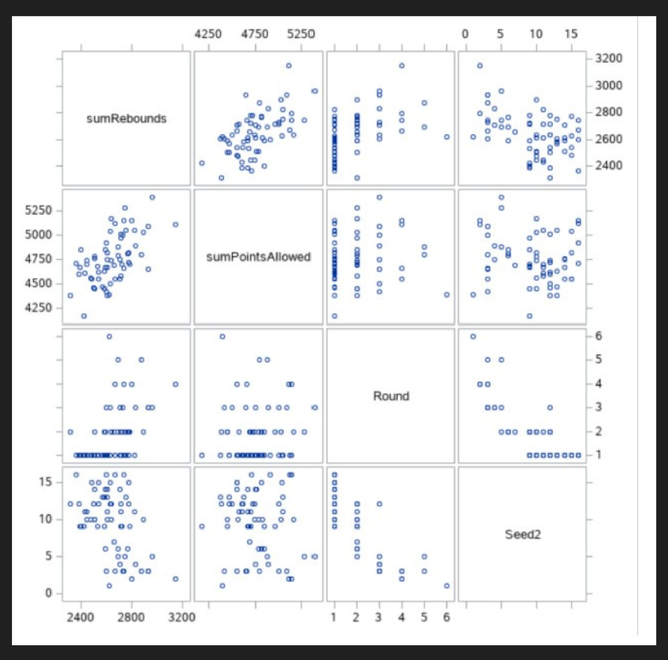
Final Model

Final Overall: $\text{GameTotal} = \beta_0 + \beta_1\text{sumRebounds} + \beta_2\text{sumPointsAllowed} + \beta_3\text{Round} + \beta_4\text{Seed2}$

FINAL MODEL STATISTICS

Adj. R-Squared: .4897
P-value: <.0001
Root MSE: 12.207

Multicollinearity



Variable Screening

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	sumPointsAllowed		1	0.2537	0.2537	8.4901	20.74	<.0001
2	sumRebounds		2	0.0507	0.3044	5.9048	4.37	0.0407
3	Seed2		3	0.0420	0.3464	4.1104	3.79	0.0564

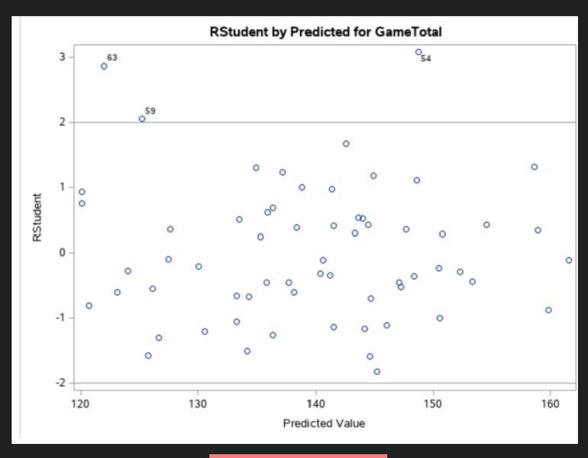
Stepwise Selection

- Resulted in sumPointsAllowed, sumRebounds, and Seed2
- All significant parameters used in final model
- SLEntry and SLStay values of 0.15

Multicollinearity Check

- VIF numbers are all below 10
- Highest correlation is .57 between rebounds and opponent total.
- Overall multicollinearity shouldn't be an issue.

Influential Observations



1. 54, 59, and 63 are outliers and were removed

Conclusions:

Prediction Equation:

$\widehat{GameTotal} = -14.93 - .022sumRebounds + .05sumPointsAllowed - 8.72Round - 1.62Seed2$

Interpretation:

The Round that the game is the largest parameter and has a lot weight in predicting the total score. The strongest positive parameter was that of the sumPointsAllowed

Usefulness for Predictions:

- the model is significant and somewhat useful for predicting the total number of points in a given March Madness game
- the R-squared value is only .51, meaning that only 51% of the variation in the data can be explained by the model

Examples of Model Usage:

UVA vs. Ohio, first round game from 2021 tournament
Over/Under: 130
Actual Total Points: 120
Predicted Total Points: 118.92
Residual: 1.08
Low residual of 1.08 and the model also correctly predicted the under for this game (the application of our model)

Future Use/Limitations:

The scope of the data is only from the 2019 tournament. With more years of data, like from the 2000s for example, the model's accuracy may be improved.
The game constantly changes in terms of strategies, so having this year's data from the 2021 tournament would also likely improve the model.
We could also improve this research by using more advanced statistics, like offensive and defensive efficiency or offensive and defensive ratings.

WORKS CITED

Blinder, Alan. “The N.C.A.A. Tournament's Pandemic Playbook.” The New York Times, The New York Times, 15 Mar. 2021, www.nytimes.com/2021/03/15/sports/ncaabasketball/ncaa-covid-march-madness.html.

Chiusano, Anthony. “How Free Throw Percentages Factor into March Madness Success.” NCAA.com, NCAA.com, 3 Aug. 2018, www.ncaa.com/news/basketball-men/article/2018-08-02/how-team-free-throw-percentages-factor-march-madness-success.

Steinberg, Leigh. “‘March Madness’ Popularity Off The Charts.” Forbes, Forbes Magazine, 17 Mar. 2016, www.forbes.com/sites/leighsteinberg/2016/03/16/march-madness-popularity-off-the-charts/?sh=14a7651b2084.