BUSOBA 7334
Sports Analytics – Final Project

Game Excitement Index:
What Makes a March Madness Game Exciting?

Nick Faupel

Date: July 8, 2023

Introduction

The NCAA Basketball Tournament is one of the most exciting and anticipated sporting events every year. Each year, the 68 best teams in college basketball participate in a single-elimination knockout tournament to determine a national champion. The tournament is often referred to as simply "March Madness" due to the amount of exciting basketball and upsets that occur each year. There are a total of 67 games played in each tournament (including the last-four-in games), and the primary goal of this analysis was to determine which games were the most exciting and determine what factors may lead to a more exciting NCAA tournament game.

Game Excitement Index (GEI) is a measure that has been developed by several different analysts for different sports. The calculation attempts to use the running win probability of a sporting event to quantify how exciting the game was. Luke Benz created a function within his ncaahoopR package that can calculate GEI given an ESPN game ID (GEI: An In-Depth Exploration, Luke Benz). The formula for this calculation can be seen below:

$$GEI = \frac{2400}{t} \sum_{i=2}^{n} |p_i - p_{i-1}|$$

After game IDs were sourced and GEI was calculated for the previous nine NCAA tournament games, further analysis was performed to determine what variables may be used to predict game excitement.

Data Overview

Prior to calculating the GEI for each NCAA tournament game, the unique ESPN game ID's needed to be pulled. This was done by first creating a list of the 68 participants for the previous nine NCAA tournaments and cleaning it up in a .csv file. The get_schedule() function from the ncaahoopR package was then nested in a for loop and used to scrape the corresponding season schedule for each of the tournament participants that year. The season data was then filtered to only contain the range of games that occurred during the duration of that year's tournament, which produced a dataset containing all 67 tournament games. The get_schedule() function was not working for years prior to 2014, and the 2020 tournament was canceled due to COVID-19, which is the reason only the last nine tournaments were used for analysis.

Once the 603 unique game ID's were scraped, the next task was to calculate the GEI and pull in other variables for analysis. The separate datasets for each tournament were combined, and the game_excitement_index() function was nested inside a for loop to calculate the GEI for each tournament game. Other variables were then scraped using the get_pbp_game() function and aggregating the variables to store in the master dataframe.

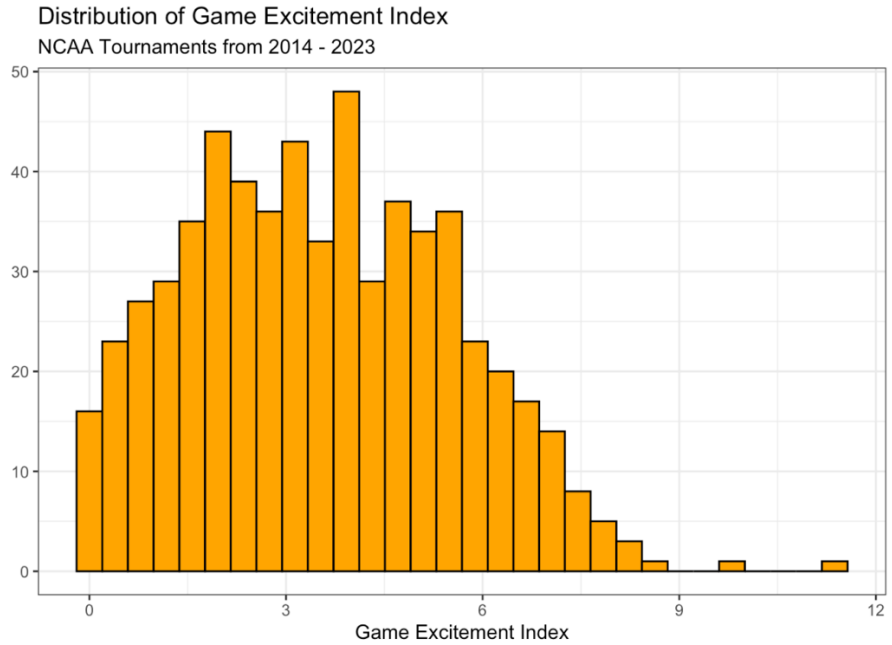
When pulling the game data from ESPN using the get_schedule() function, there was not much information included. Most of the variables were generated using aggregation with the get_pbp_game() function. Other data was scraped from Wikipedia. The following variables were collected and used for modeling purposes, with the hope to determine which variables will best predict GEI.

Variable	Description
FT	Number of Free Throw Attempts
Threes	Number of Three Point Attempts
Play Length	Average length of plays throughout the game
Away Seed	Away Team Seed
Home Seed	Home Team Seed
Round	Round of Tournament
Score Diff End	Score Difference at End of Game
Total Score	Total Points Scored
Outcome	Result was Upset or Expected

A full data dictionary can be found in the appendix at the end of this report.

Exploratory Analysis

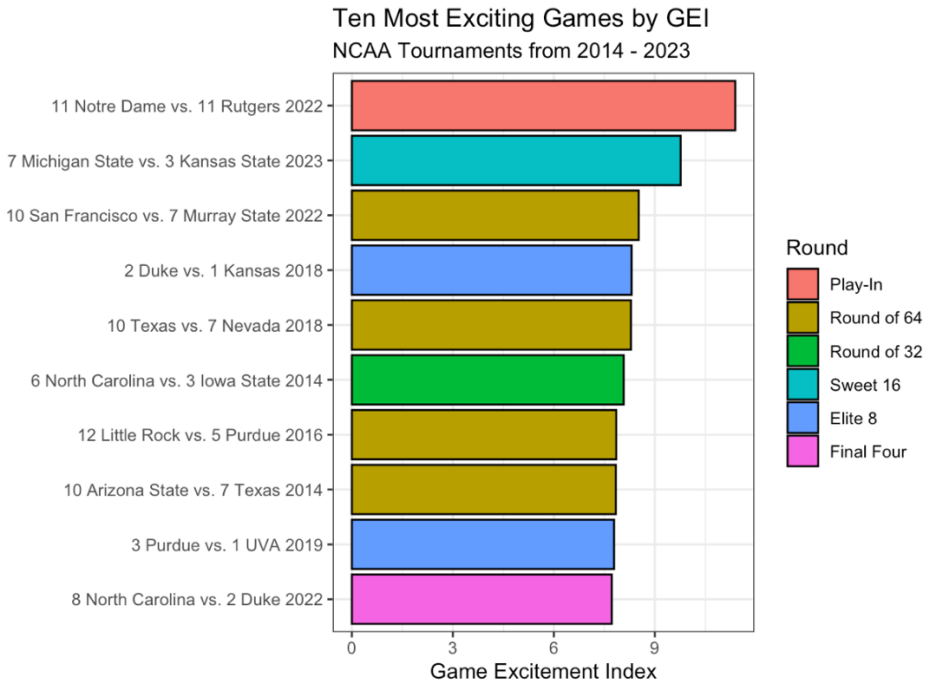
First, we created a histogram of GEI to gain a better understanding of the distribution among the games in our dataset. Most games fall between 0 and 6, while there are a couple outliers that have a GEI higher than 9.

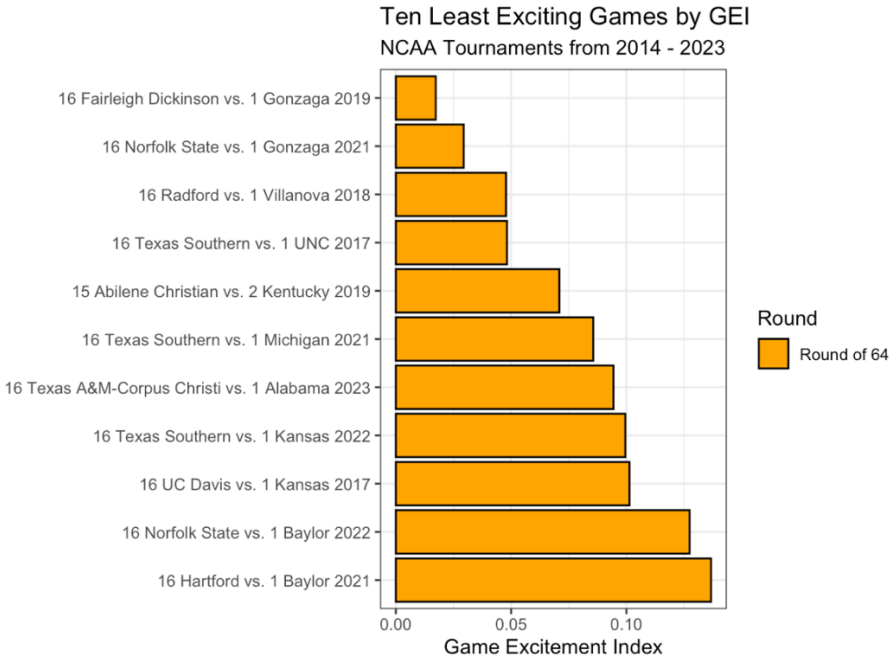


An overview of the top 10 games by GEI can be seen below with some game summary statistics included as well. The most exciting game in our dataset was a play-in game played between Notre Dame and Rutgers in 2022 in which Notre Dame won by 2 points in overtime. While the betting line for this game was evenly split, Rutgers technically had a higher overall rank heading into the game, which is why it is listed as an upset.

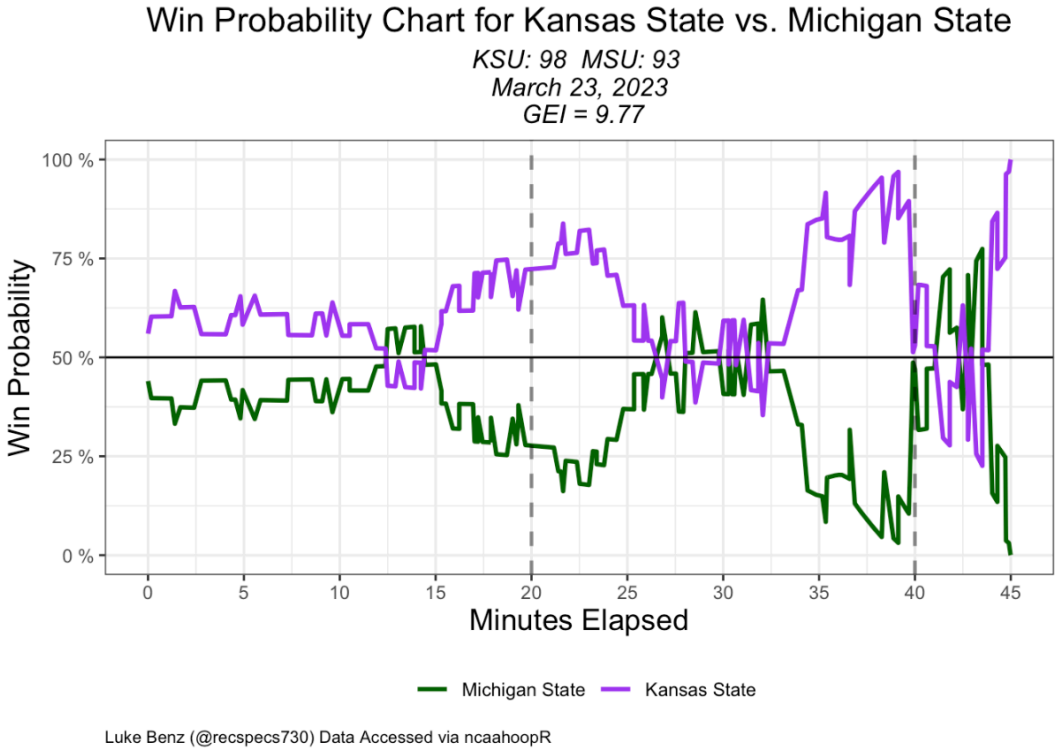
Matchup	Round	Winner	Outcome	MeanScoreDiff	AwayScore	HomeScore	GEI
11 Notre Dame vs. 11 Rutgers 2022	Play-In	Notre Dame	Upset	0.77	89	87	11.3894
7 Michigan State vs. 3 Kansas State 2023	Sweet 16	Kansas State	Expected	1.93	93	98	9.7685
10 San Francisco vs. 7 Murray State 2022	Round of 64	Murray State	Expected	1.10	87	92	8.5219
2 Duke vs. 1 Kansas 2018	Elite 8	Kansas	Expected	0.29	81	85	8.3098
10 Texas vs. 7 Nevada 2018	Round of 64	Nevada	Expected	4.30	83	87	8.2920
6 North Carolina vs. 3 Iowa State 2014	Round of 32	Iowa State	Expected	0.42	83	85	8.0754
12 Little Rock vs. 5 Purdue 2016	Round of 64	Little Rock	Upset	3.02	85	83	7.8548
10 Arizona State vs. 7 Texas 2014	Round of 64	Texas	Expected	4.94	85	87	7.8426
3 Purdue vs. 1 UVA 2019	Elite 8	UVA	Expected	0.21	75	80	7.7885
8 North Carolina vs. 2 Duke 2022	Final Four	North Carolina	Upset	0.06	81	77	7.7196

The following two graphs show the top 10 and bottom 10 games according to GEI. The top 10 games occurred in several different rounds of the tournament, while the bottom 10 games all occurred in the round of 64. The bottom 10 games were, unsurprisingly, mostly number one seeds blowing out 16 seeds.

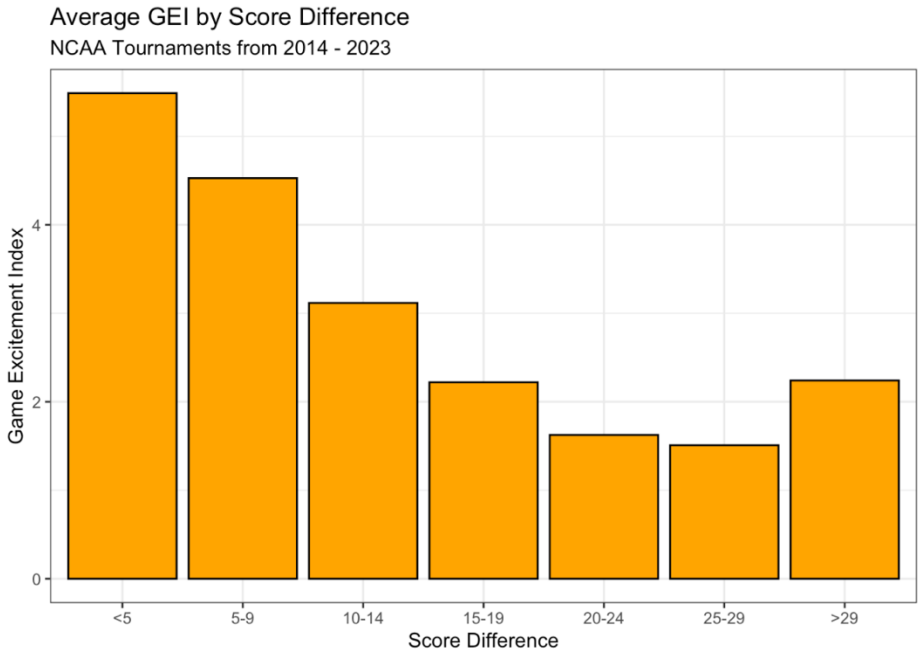




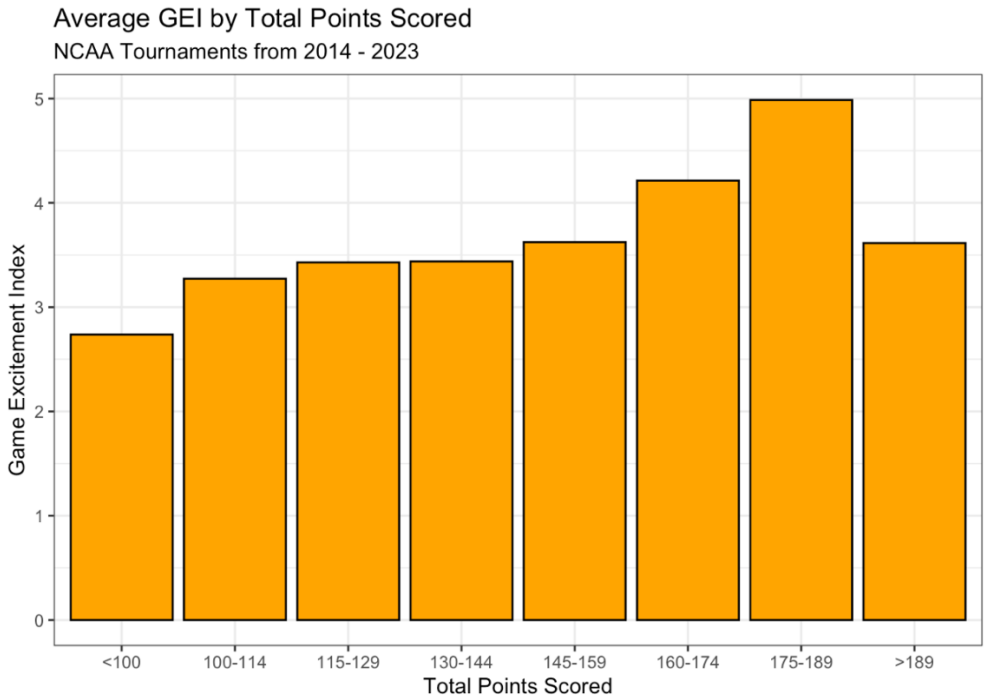
The following graph shows the ESPN Win Probability chart for the game between Kansas State and Michigan State in the Sweet 16 of this year's (2023) March Madness tournament. This game was second overall in our dataset, having a GEI of 9.77. This game went to overtime and the two teams made exciting play after exciting play all game on both ends of the floor. Commentators said several times that the game was an instant classic.



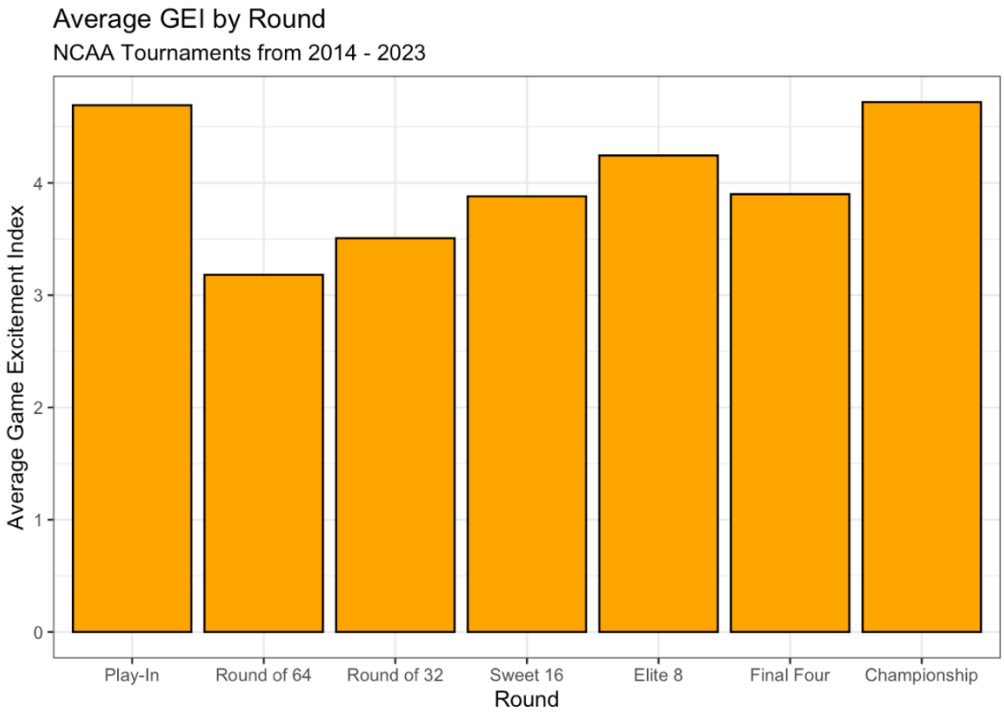
We next looked at how points scored and point differentials impacted GEI. The following graph shows that games ending with a score differential less than 5 are by far the most exciting. As the score difference increases, GEI tends to decrease. Interestingly, average GEI was higher when the end score differential is greater than 29 points, which does not fit the trend. Winning by 30 or more points would indicate a complete blowout, which are typically not very exciting games. However, these games did contain some big upsets which could be a reason for the uptick in average GEI.



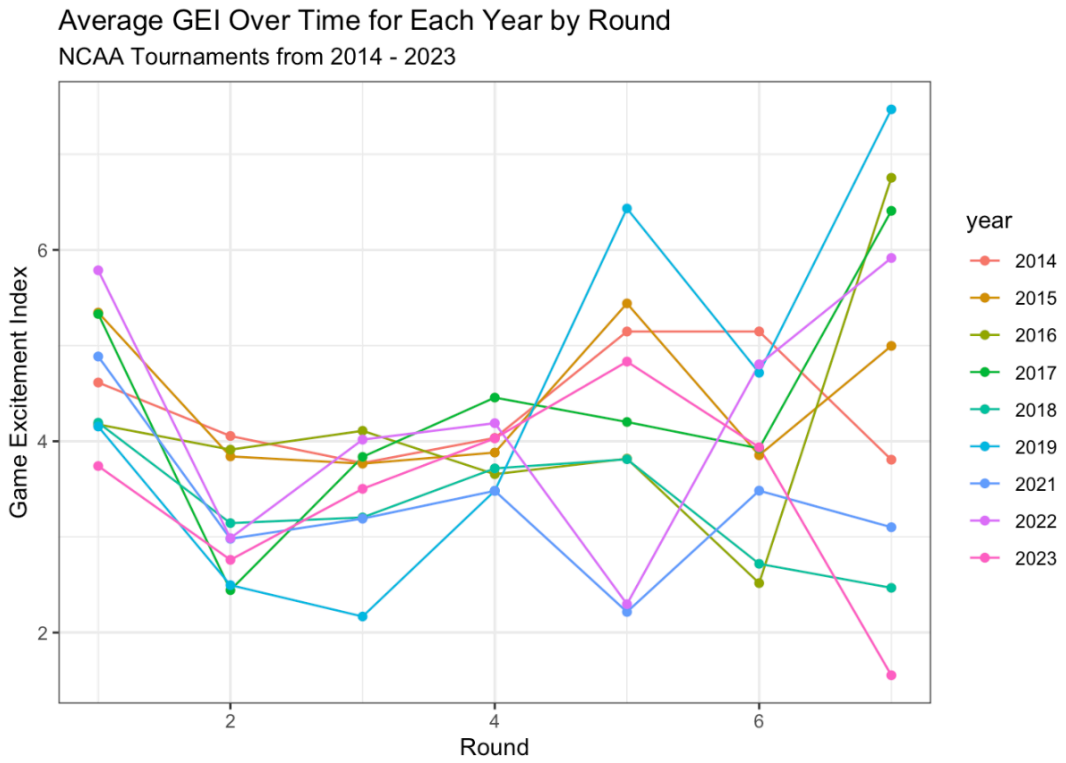
This next graph shows how GEI tends to increase as more points are scored.



Finally, we were also curious about how GEI fluctuates from round to round in the tournament. Play-In games tend to have a higher GEI than most other rounds, often on par with championship games. This is likely due to the even match between the teams. This graph also shows that Elite 8 games tend to be more exciting than the following Final Four games. Teams have more time to prepare for their Final Four matchups than they do for their Elite 8 games, which could be a reason why there is less excitement. With less time to prepare, coaches and players often have to make more in-game adjustments, which can result in closer games, more scoring runs, lead changes, and set the stage for more upsets.

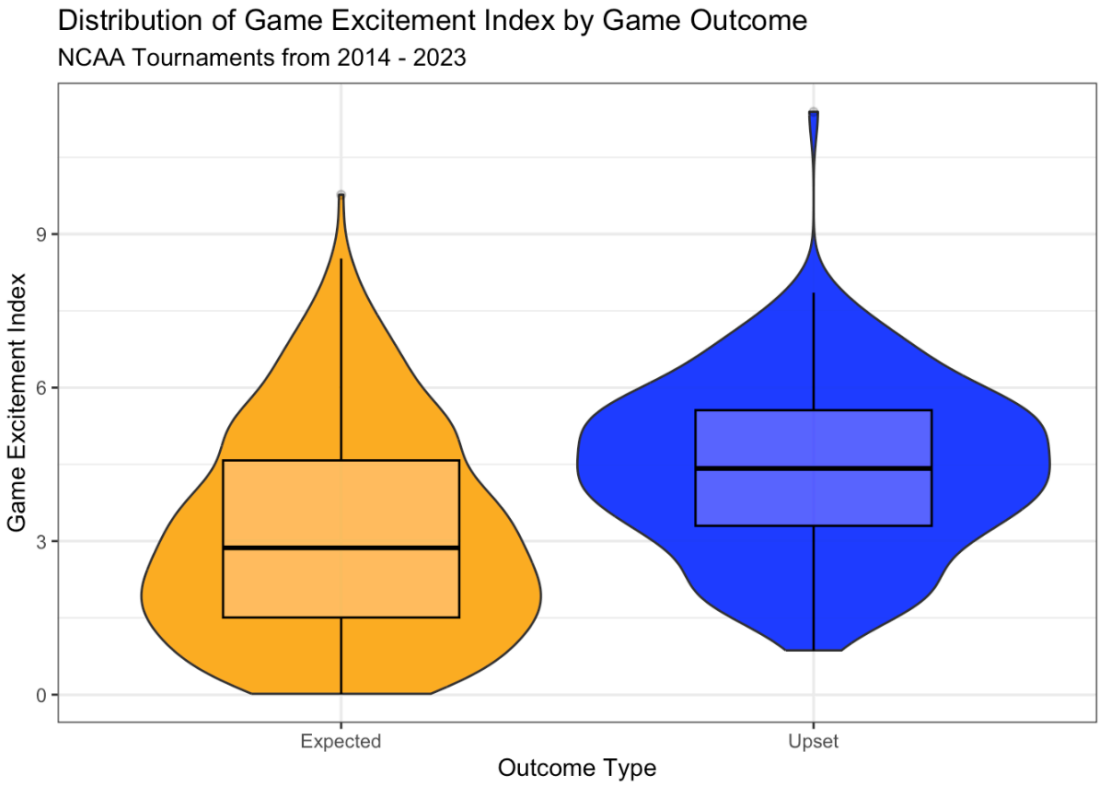


When looking at GEI variation by round over the years, we see that championship games can either be very exciting or not as exciting, which was the case with the most recent championship matchup between UConn and San Diego State.

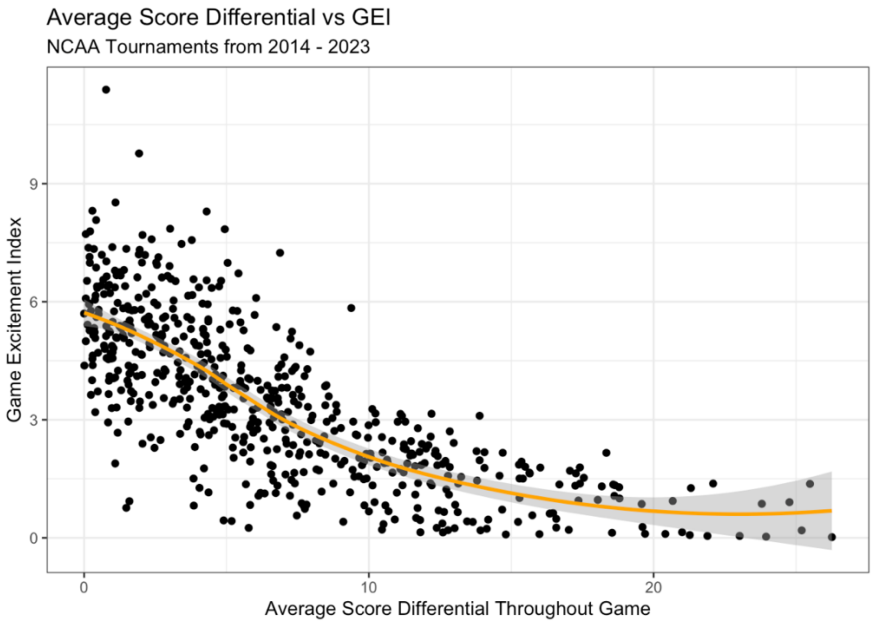


Further Analysis

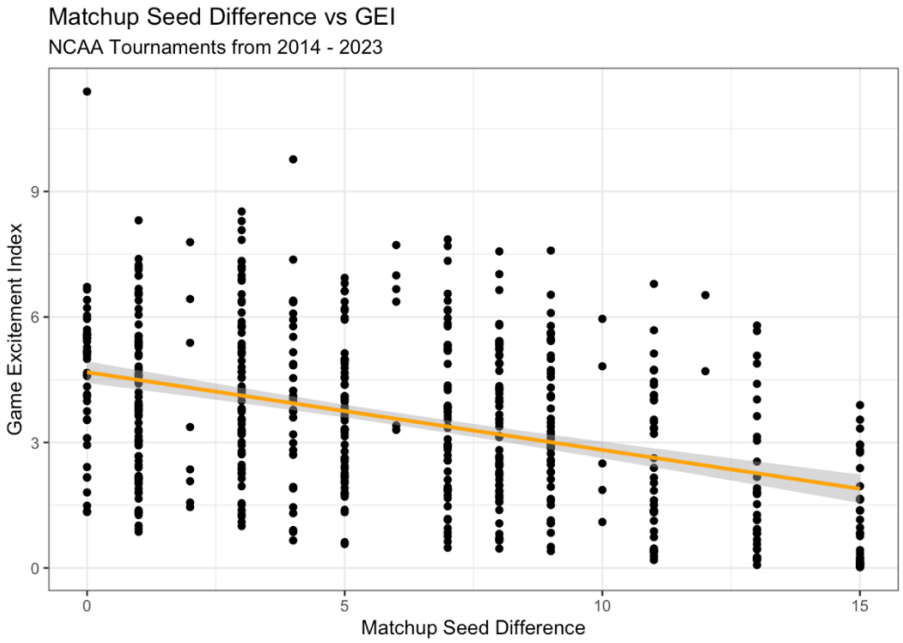
To gain a better understanding of GEI and influential factors that influence GEI, we turned to statistical methods using our dataset. We wanted to determine what variables in our dataset proved to be most related to GEI. Following our EDA, we had an inclination that upsets were more likely to lead to a higher GEI compared to games in which the expected winner did in fact win the game. The following violin graph shows that this is indeed correct. In general, upsets have higher GEIs, having a median GEI around 4, while games in which the expected outcome occurs have a median GEI of just below 3.



We also figured that games in which the average score differential throughout the game was lower most likely tended to have higher GEIs on average. This also proved true and can be seen in the following graph. Closer games are usually much more exciting than blowouts.



Another variable that shared a relationship with GEI was the seed difference of the matchups. The following graph shows how games played between two teams with similar seeds tend to be more exciting. A seed difference of 0 would indicate a game being played by teams of the same seed. This is only possible in play-in games and in the Final Four and Championship games. As such, it makes sense that these games are much more exciting on average compared to the 1 vs 16 seed games that happen in the round of 64.



Finally, we employed a linear regression model to determine which variables in our dataset had the best ability of explaining the variance in GEI. For our model, we predicted GEI using free throw attempts, three-point attempts, average score differential throughout the game, average play length, both team's seeds, the round of the tournament, the score difference at the end of the game, total points scored, and whether the outcome of the game was an upset or expected.

Results for our model are as follows:

```
#Fit a regression with all variables
final_model <- lm(gei ~ FT + threes + avg_score_diff + play_length + away_seed + home_seed + round_name + score_di
ff_end + tot_score + outcome, data = df)

summary(final_model)
```

```
##
## Call:
## lm(formula = gei ~ FT + threes + avg_score_diff + play_length +
##     away_seed + home_seed + round_name + score_diff_end + tot_score +
##     outcome, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0081 -0.7849 -0.0944  0.6977  4.1235
##
## Coefficients:
##                          Estimate Std. Error t value             Pr(>|t|)
## (Intercept)              0.007327   1.203216   0.006              0.99514
## FT                       0.012021   0.006019   1.997              0.04628 *
## threes                  -0.006657   0.006363  -1.046              0.29594
## avg_score_diff          -0.259426   0.010467 -24.785 < 0.0000000000000002 ***
## play_length              0.238046   0.074750   3.185              0.00153 **
## away_seed2               0.860531   0.629442   1.367              0.17213
## away_seed3              -0.225989   0.578955  -0.390              0.69643
## away_seed4              -0.518862   0.572172  -0.907              0.36489
## away_seed5              -0.376558   0.572752  -0.657              0.51116
## away_seed6              -0.026642   0.598763  -0.044              0.96453
## away_seed7              -0.366579   0.563737  -0.650              0.51579
## away_seed8               0.403793   0.548757   0.736              0.46214
## away_seed9              -0.490620   0.572619  -0.857              0.39192
## away_seed10             -0.042302   0.577276  -0.073              0.94161
## away_seed11             -0.502051   0.563004  -0.892              0.37292
## away_seed12             -0.600075   0.639414  -0.938              0.34840
## away_seed13             -0.640757   0.630773  -1.016              0.31015
## away_seed14             -1.404409   0.609917  -2.303              0.02167 *
## away_seed15             -1.038285   0.582674  -1.782              0.07530 .
## away_seed16             -0.705120   0.598215  -1.179              0.23902
## home_seed2               0.236050   0.240873   0.980              0.32752
## home_seed3               0.862898   0.285306   3.024              0.00260 **
## home_seed4               0.478771   0.262882   1.821              0.06910 .
## home_seed5               0.660995   0.332442   1.988              0.04726 *
## home_seed6               0.763777   0.353045   2.163              0.03093 *
## home_seed7               0.715426   0.354490   2.018              0.04405 *
## home_seed8               0.903265   0.333211   2.711              0.00692 **
## home_seed9              -0.194937   0.609387  -0.320              0.74917
## home_seed10              0.932645   0.648095   1.439              0.15069
## home_seed11              0.986707   0.624080   1.581              0.11443
## home_seed12             -0.463041   0.862791  -0.537              0.59170
## home_seed15              0.067007   1.308908   0.051              0.95919
## home_seed16              0.820977   0.763398   1.075              0.28265
## round_nameRound of 64   -0.357452   0.685933  -0.521              0.60249
## round_nameRound of 32   -0.389968   0.641931  -0.607              0.54377
## round_nameSweet 16      -0.074362   0.668727  -0.111              0.91150
## round_nameElite 8       -0.094256   0.687818  -0.137              0.89105
## round_nameFinal Four    -0.417703   0.702556  -0.595              0.55239
## round_nameChampionship   0.171384   0.849274   0.202              0.84015
## score_diff_end           0.007940   0.006072   1.308              0.19155
## tot_score                0.023383   0.003083   7.585    0.000000000000139 ***
## outcomeUpset             0.725905   0.155234   4.676    0.000003666747730 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.165 on 560 degrees of freedom
## Multiple R-squared:  0.6867, Adjusted R-squared:  0.6638
## F-statistic: 29.94 on 41 and 560 DF,  p-value: < 0.00000000000000022
```

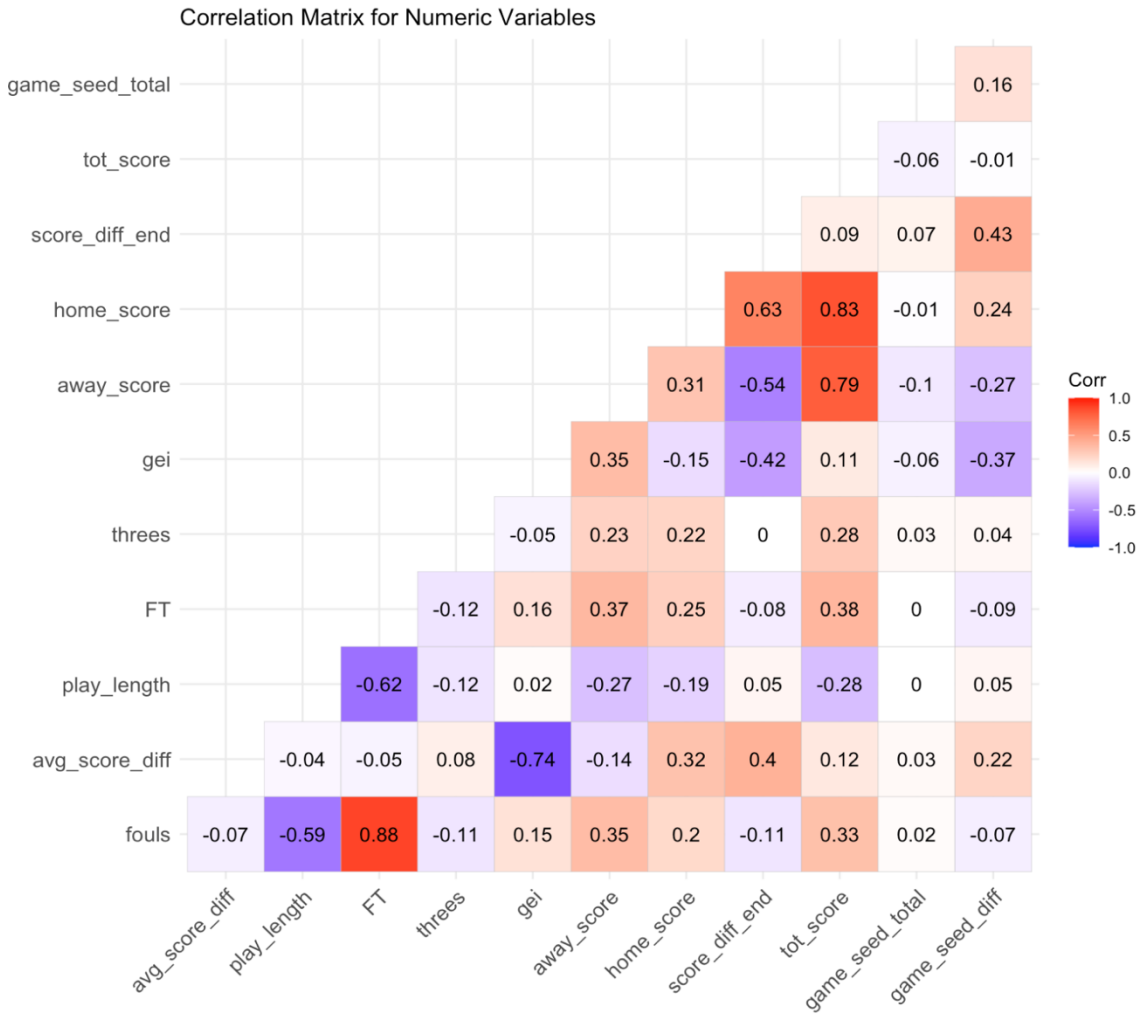
Our model had an adjusted R^2 of 0.6638. The predictor variables with the lowest p-values, and therefore the most statistical significance, were average point differential, total points scored, and outcome being an upset, away seed and home seed, and the average length of plays. We felt all of these made sense given our understanding of basketball and our findings in the EDA. It was interesting to find that the round of the tournament did not serve as a good indicator of GEI, but this is not necessarily surprising given the variability that occurs from round to round. After all, it is called March Madness for a reason.

Conclusion

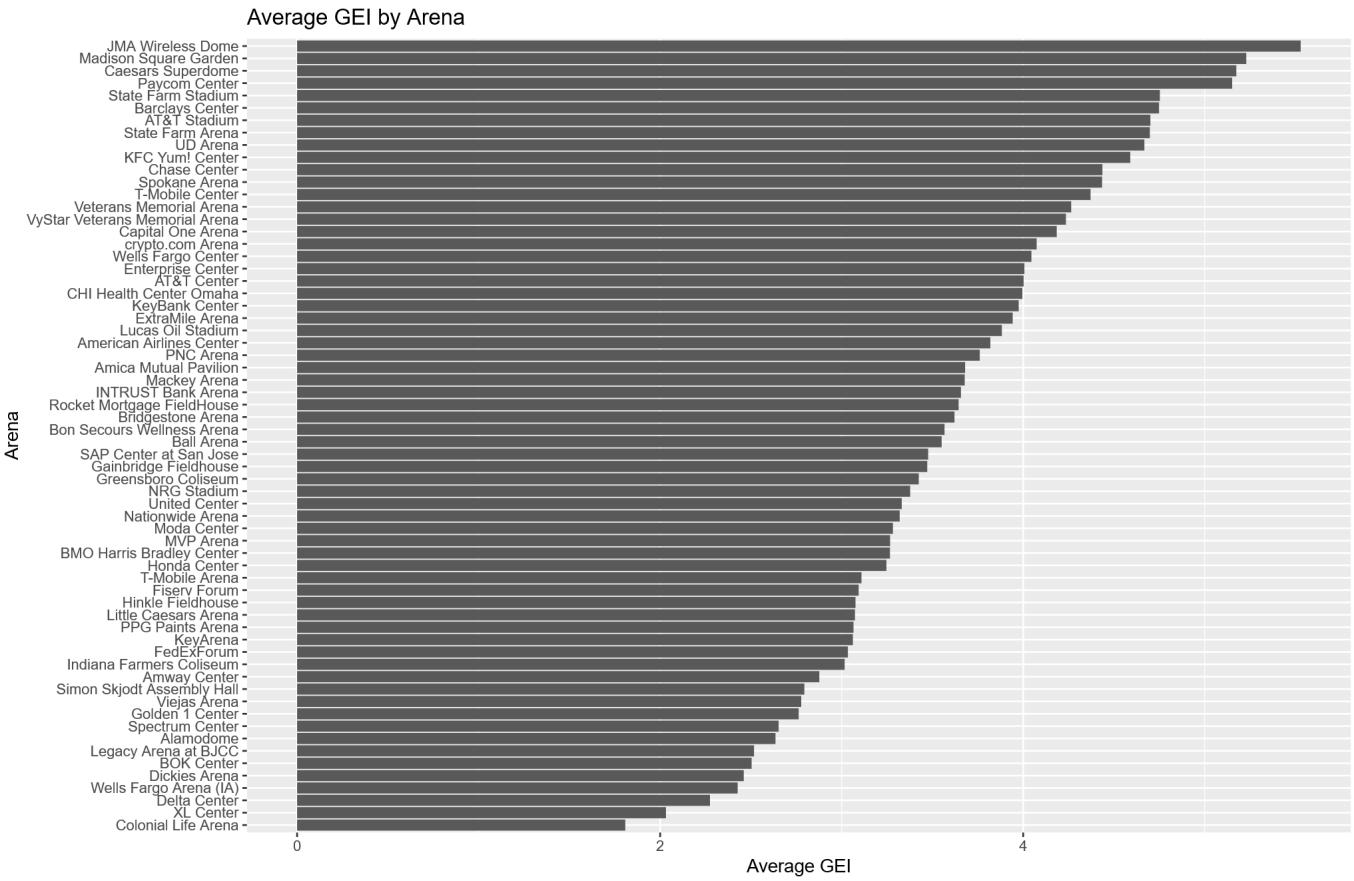
This analysis of the NCAA Basketball Tournament aimed to determine the factors that contribute to the Game Excitement Index. By calculating the GEI for previous tournament games and exploring various variables, we gained several insights. Our findings revealed that games with smaller score differentials, closer seeds, and upsets tend to have higher GEIs. Games with more points scored were generally also more exciting. The round of the tournament did not consistently indicate higher excitement levels, as Play-In games often rivaled championship games in terms of excitement. A linear regression model was utilized to predict GEI, with average point differential, total points scored, and game outcome (upset or expected) proving to be the most significant predictors. Overall, this analysis enhanced our understanding of the factors that contribute to an exciting NCAA tournament game, highlighting the unpredictable and thrilling nature of March Madness.

Appendix

Correlation plot of numeric variables



We also analyzed average GEI by arena. The JMA Wireless Dome had the most exciting games.



Data Dictionary:

Field	Definition
game_id	Game ID
matchup	Teams Playing
date	Date of Game
tourney_year	Tournament Year
ref_crew	Referee Crew
arena	Arena
fouls	Number of Fouls
avg_score_diff	Average Score Differential During Game
play_length	Average Play Length in Seconds
FT	Number of Free Throw Attempts
threes	Number of Three Point Attempts
away_team	Away Team
away_conf	Away Team Conference
away_berth	Away Team Tournament Bid Type
away_seed	Away Team Seed
away_rank	Away Team Rank
away_record	Away Team Regular Season Record
home_team	Home Team
home_conf	Home Team Conference
home_berth	Home Team Tournament Bid Type
home_seed	Home Team Seed
home_rank	Home Team Rank
home_record	Home Team Regular Season Record
gei	Game Excitement Index
round	Round of Tournament expressed as a Number
round_name	Round of Tournament
away_score	Away Team Score
home_score	Home Team Score
score_diff_end	Score Difference at End of Game
tot_score	Total Points Scored
seed_matchup	Seed No. vs Seed No.
game_seed_total	Sum of Team Seeds
game_seed_diff	Difference Between Home Team Seed and Away Team Seed
winner	Winning Team
outcome	Result was Upset or Expected