INFO 526 Summer 2025

# FINAL PROJECT:
# NCAA March Madness Analysis

By Matt Osterhoudt

# Topic and Motivation

- The NCAA March Madness is an annual single-elimination basketball tournament consisting of 68 teams from the NCAA Division 1 leagues.
- It is one of the biggest annual sporting events in the US, known for its intense competition and dramatic upsets.
- As of last year, 3.1 billion dollars in bets were placed on March Madness.

- My motivation for analyzing some of the March Madness data stems from my interest in basketball. It's also a fun annual sporting event driven by national competitive spirit for basketball fans everywhere.
- Connecting basketball to data science/analytics.
- Develop insights on tournament data: useful for predictions and (less ethically) placing sports bet.

# The Data: Preview

**Team_results:**

- PASE: Performance Against Seed Expectations
- F4PERCENT: Likelihood of a team getting to at least 1 Final Four
- R64: Amount of times the team made it to the Round of 64

**Matchups_data:**

- winning_team_seed: Seed number of winning team
- losing_team_seed: Seed number of losing team
- round_of: Teams left per round

**Public_picks:**

- F4: The percent of people who picked the team to win the game in the Final 4

TidyTuesday

Source: Kaggle

GitHub

Source: Shoenot

# Main Questions:

Q1: How well does past tournament performance from 2008-2023 correlate with predictions for the 2024 tournament?

Hypothesis: There is little to no correlation between past tournament performance and predictive behavior.

Q2: Are lower-seeded teams winning more frequently over time, and if so, by what level of magnitude?

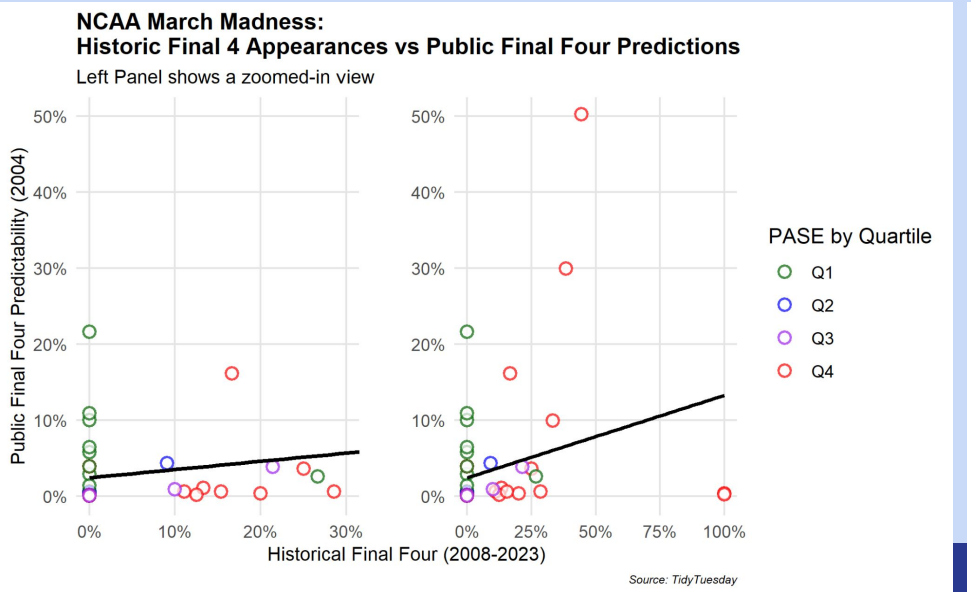
Hypothesis: Over time, there have been more upsets with an increasing level of magnitude.

# Visualization: Question 1

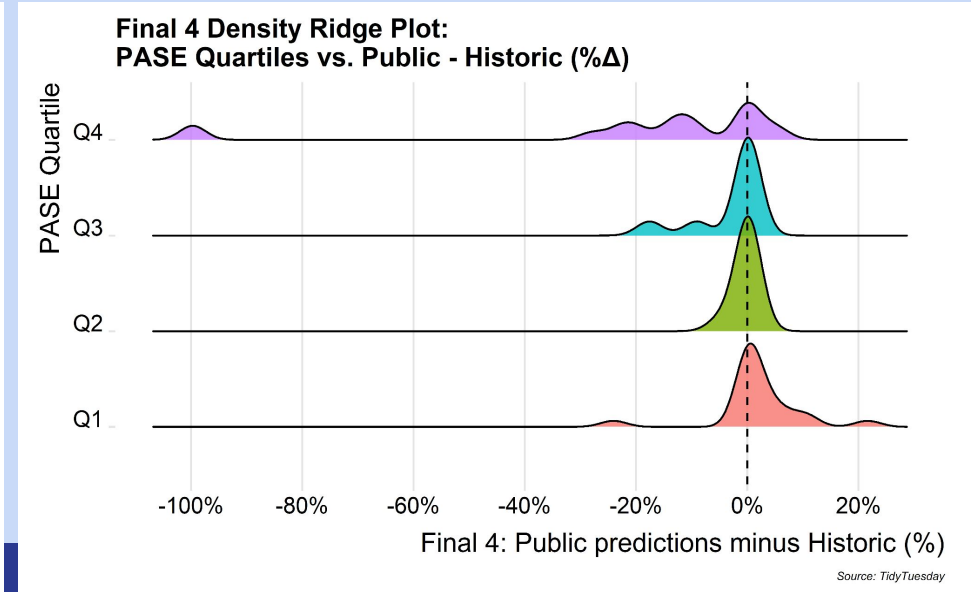
Q1: How well does past tournament performance from 2008-2023 correlate with predictions for the 2024 tournament?

Hypothesis: There is little to no correlation between past tournament performance and predictive behavior.

## Scatter Plot:




## Density Ridge Plot:

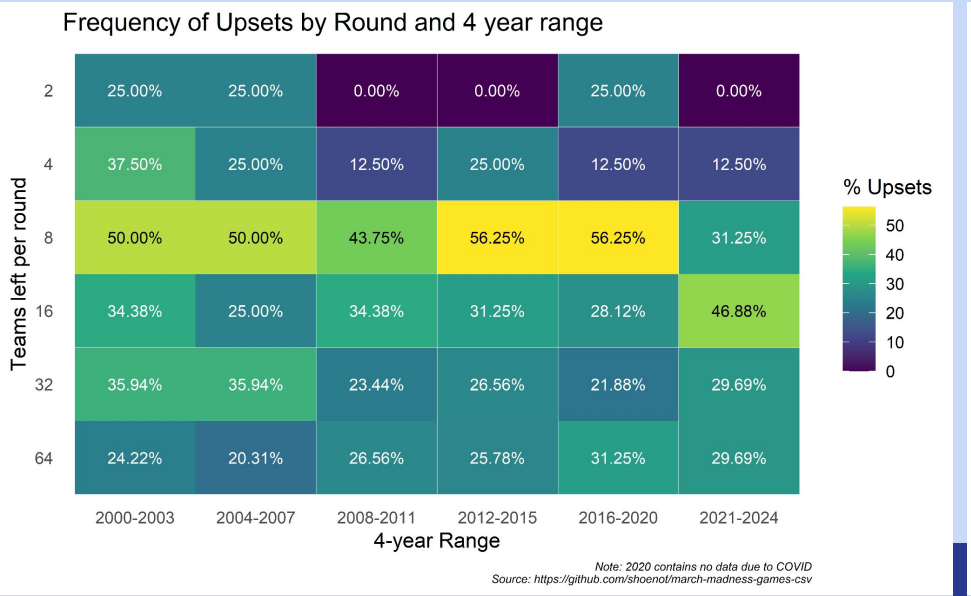


# Visualization: Question 2

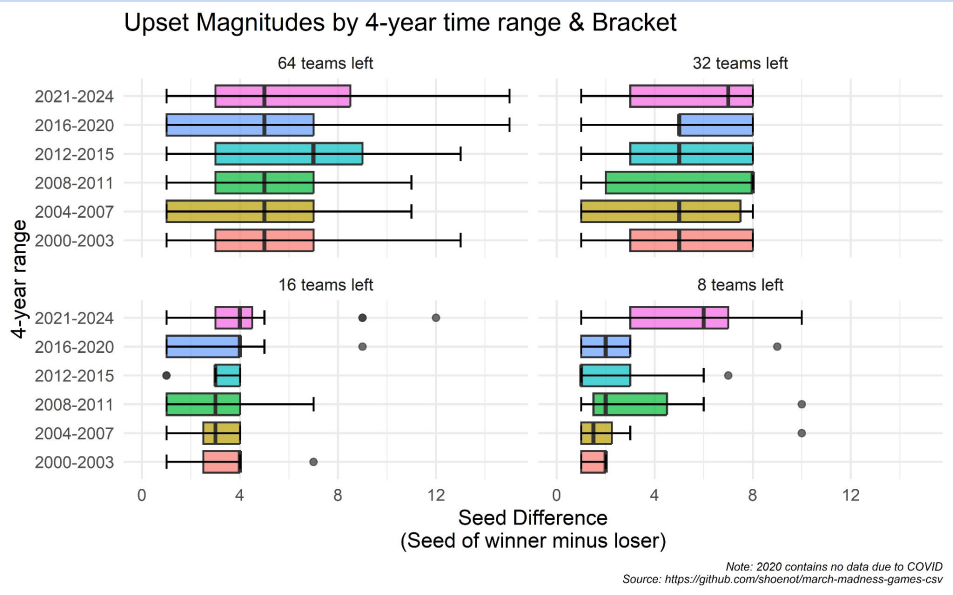
Q2: Are lower-seeded teams winning more frequently over time, and if so, by what level of magnitude?

Hypothesis: Over time, there have been an more upsets with an increasing level of magnitude.

## Heat Map Tile Plot:




## Box and Whisker Plot:



# Conclusions:

## Question 1:

- Cannot conclusively confirm if past performance correlates with future performance
- Positive relationship appears to be established, but doubtful. Other factors likely at play
- Would require more data or better usage of dataset

## Question 2:

- I can conclude there appears to be a higher frequency of upset games AND with stronger magnitude over time
- More evident at earlier rounds compared to later rounds
- Both box and whisker plot and heatmap plots demonstrated this pretty well

# Future Work:

- Consider using additional data. I felt limited to the dataset I had, and the first question could not be done over time. For example, adding data based on team performances over a season.
- For the second question, I could also analyze the point differential per match. It was another unexplored variable.
- Using a higher level of statistical modeling to interpret trends and analysis.