METHOD TO THE MADNESS

Analyzing Trends and Predictive Insights in
the NCAA Men's Basketball Tournament

INFO 526 – Summer 2025
Reagan Fleming

INTRODUCTION AND MOTIVATION

- **Ask any College Basketball fan what they love about the NCAA tournament and you will likely receive 2 answers.**
  - **Filling out their bracket, attempting to predict the unpredictable.**

  - **The upsets, surprises, and Cinderella stories that inevitably occur every year.**

- **This project aims to:**
  - **quantify the factors that inspire fan predictions**

  - **Identify overperforming and underperforming teams**

  - **Identify upset patterns that have occurred throughout tournament history**
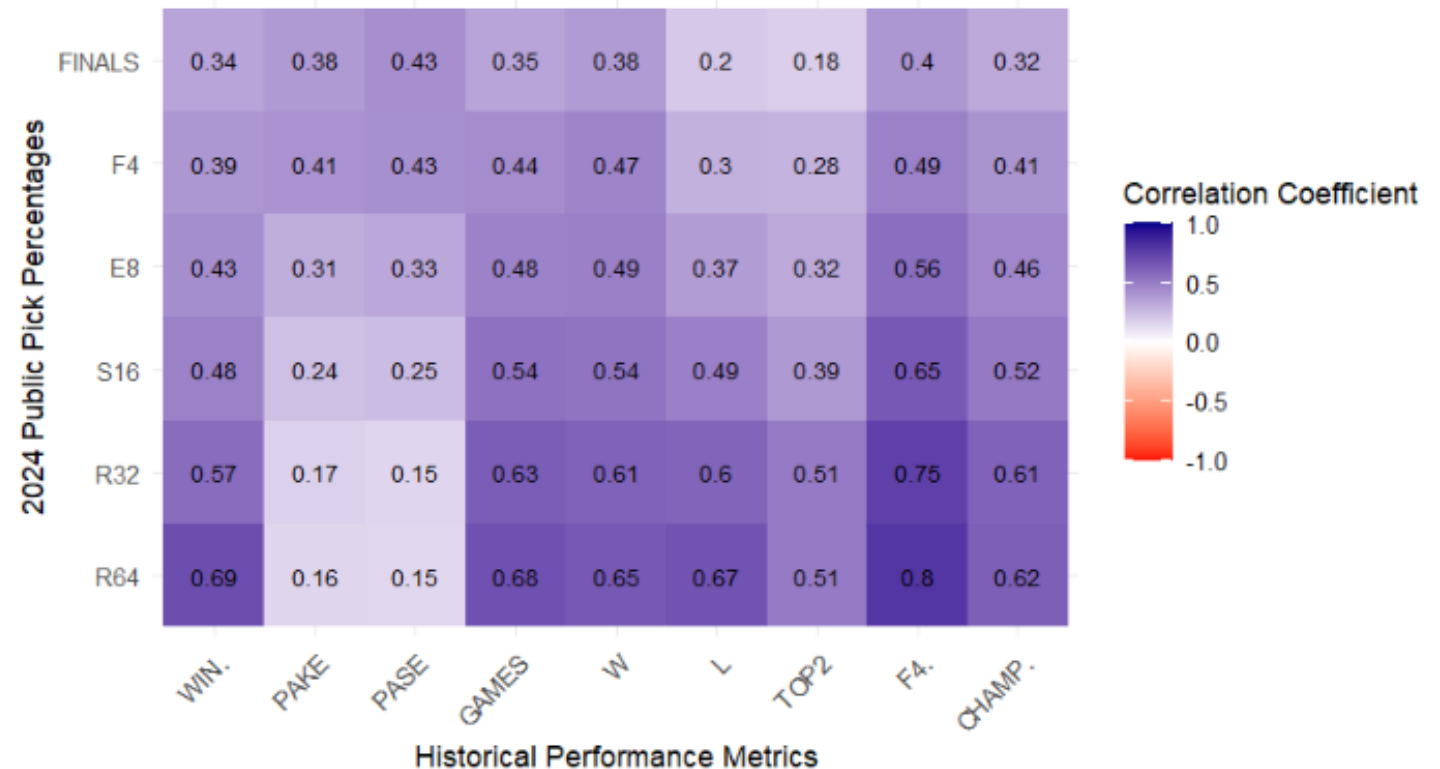
INTRODUCING THE DATA

- This analysis used data from 2 sources
  - MarchMadnessDatasets – a github repo from user christoukmaji containing win/loss data from all tournaments from 2011-2022
    - [christoukmaji/MarchMadnessDatasets: March Madness Datasets from 2002-2019, 2021. Datasets curated by Chris Toukmaji, data published by Ken Pomeroy. (github.com)](github.com)

  - NCAA Men's March Madness – sourced from Nishaan Amin's Kaggle dataset. Contains additional historical tournament data as well as public tournament prediction data.
    - [tidytuesday/data/2024/2024-03-26/readme.md at master · rfordatascience/tidytuesday (github.com)](github.com)

FACTORS INFLUENCING PUBLIC PICKS

- Correlation between historical performance metrics and 2024 public pick percentages

- High correlation in early rounds (R64-S16) with winning percentage (WIN.) and final four appearances (F4.)

- In later rounds (E8-FINALS) the highest correlating historical metrics are Performance against Seed Expectations (PASE) and Final Four Percentage

- It's no surprise that teams with a high number of final four trips are treated favorably in fan predictions

- Low correlation with PASE and PAKE in early rounds, correlation strengthens in late rounds

Correlation Coefficients between Historical Performance Metrics and 2024 Public Pick Percentages

2024 Public Pick Percentages	WIN.	PAKE	PASE	GAMES	W	L	TOP2	F4.	CHAMP.
FINALS	0.34	0.38	0.43	0.35	0.38	0.2	0.18	0.4	0.32
F4	0.39	0.41	0.43	0.44	0.47	0.3	0.28	0.49	0.41
E8	0.43	0.31	0.33	0.48	0.49	0.37	0.32	0.56	0.46
S16	0.48	0.24	0.25	0.54	0.54	0.49	0.39	0.65	0.52
R32	0.57	0.17	0.15	0.63	0.61	0.6	0.51	0.75	0.61
R64	0.69	0.16	0.15	0.68	0.65	0.67	0.51	0.8	0.62

Historical Performance Metrics

Correlation Coefficient scale: 1.0 / 0.5 / 0.0 / -0.5 / -1.0

WIN. = Winning Percentage
PAKE = Performance against Komputer Expectations
PASE = Performance against Seed expectations
GAMES = number of games played in march madness
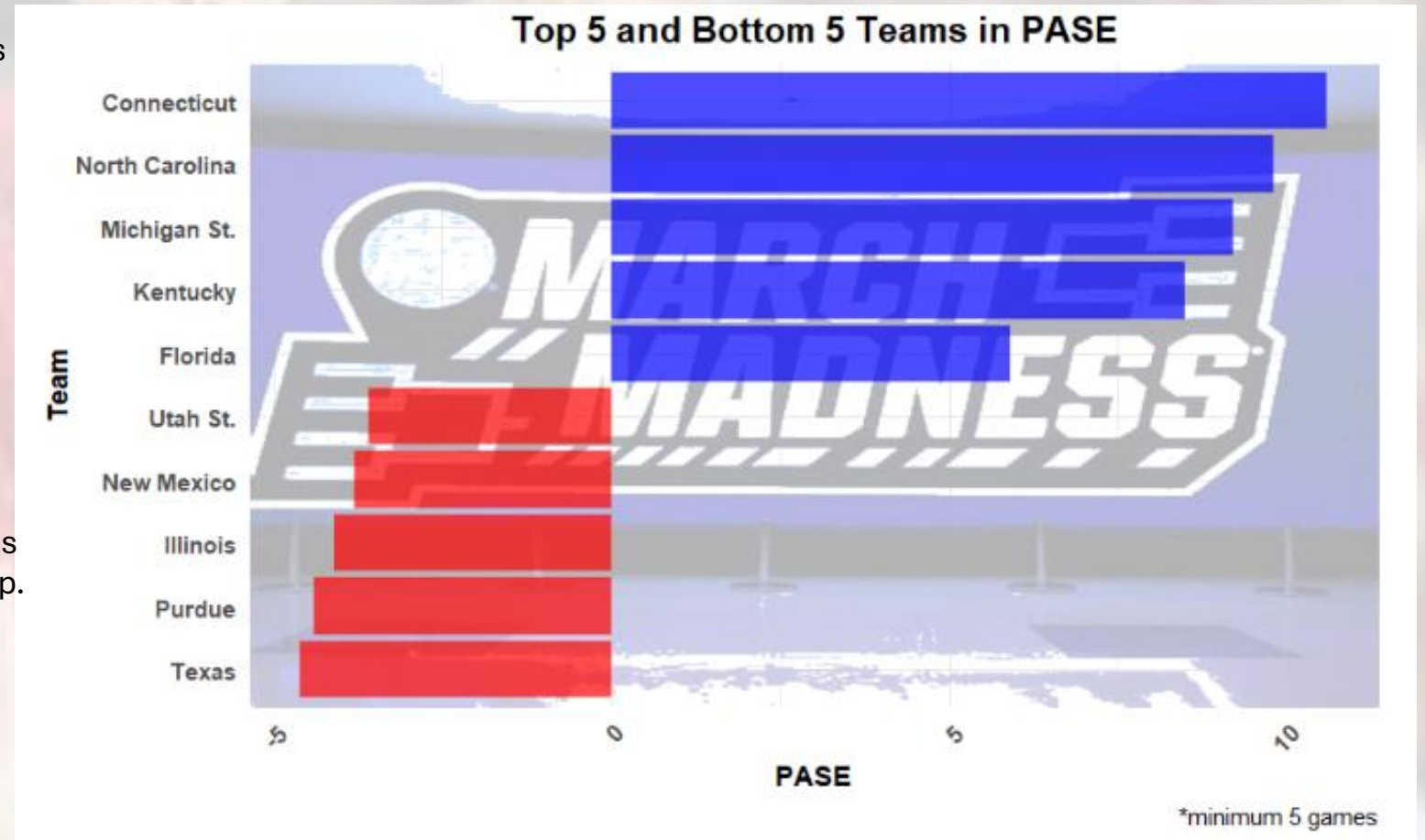W = Number of wins in march madness

TOP2 = amount of times the team has been a 1 or 2 seed
F4. = calculated likelihood of appearing in at least 1 Final 4
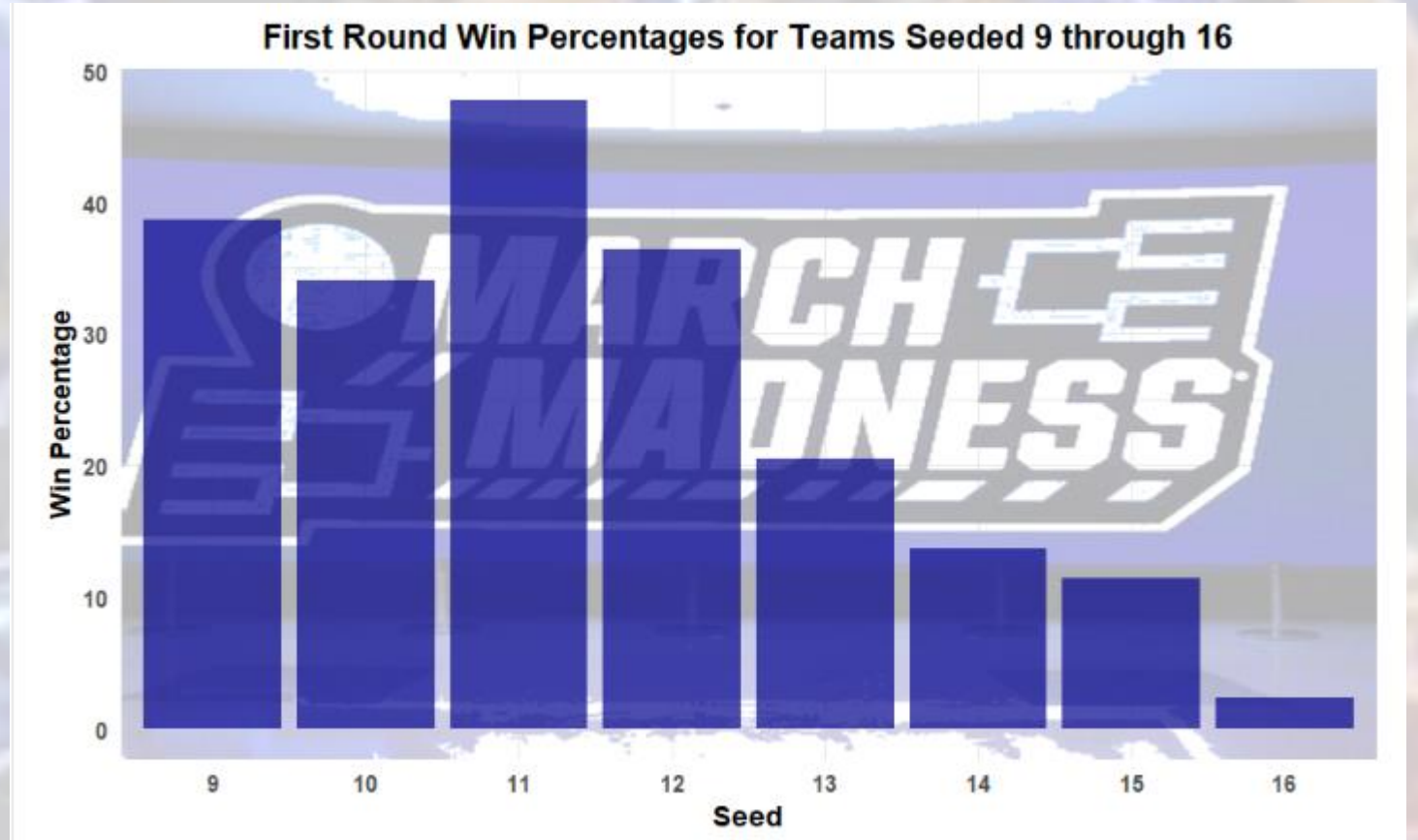CHAMP. = calculated likelihood of winning at least 1 championship

OVERPERFORMERS AND UNDERPERFORMERS

- PASE – Performance against seed expectations

- A team's seed can be seen as a metric of how many games they "should" win in the tournament

- High PASE ranking – team outperforms expectations on average

- Low PASE ranking – team underperforms expectations on average

- Last slide – PASE was the best predictor for fans to select a team to make it to the championship.

- Coincidence or no?

Top 5 and Bottom 5 Teams in PASE

Teams (top to bottom): Connecticut, North Carolina, Michigan St., Kentucky, Florida, Utah St., New Mexico, Illinois, Purdue, Texas

PASE

*minimum 5 games

FIRST ROUND UPSET FREQUENCY

- In Round 1 (R64), teams seeded 9-16 are matched with teams 8-1, respectively

- 12-5 upsets are considered very common

- Our data ranges from 2008-2024

- In that timeframe, 11-6 upsets are more common than 12-5

- Any 9-16 seeded team that defeats a 1-8 seeded team will increase their PASE score

- Utah State – Bottom 5 in PASE – has been a 12 seed 5 times since 1985
  - In those 5 games, they are 1-4

First Round Win Percentages for Teams Seeded 9 through 16

CONCLUSIONS AND FUTURE WORK

- Early rounds – fans rely on simpler metrics like Win% and previous F4 appearances

- Later rounds, PASE and PAKE play a larger role

- We see which teams over and underachieve

- 11 and 12 seeds have fared very well in recent tournaments

- There are 0 recorded instances of anyone predicting the tournament perfectly!

Future Work
- Could increased public awareness of advanced stats (PASE, PAKE) increase future prediction accuracy?

- Extend analysis to other sports (NFL, NHL, NBA playoffs)

- Interactive analysis tool or app for fans

Image sources:
- https://www.usatoday.com/picture-gallery/sports/ncaab/2014/01/18/how-the-top-25-college-basketball-teams-fared/4636123/
- https://www.ncaa.com/news/basketball-men/article/2018-10-23/duke-basketball-has-most-ap-preseason-poll-appearances
- https://www.wruf.com/headlines/2020/03/14/ncaa-basketball-tournament-cancelled/
- https://www.ncaa.com/news/basketball-men/article/2018-10-02/college-basketballs-9-winningest-teams
- https://bustingbrackets.com/2020/11/22/ncaa-basketball-power-ranking-357-teams-2020-21-season/22/