Predicting If The Shoe Fits

By: Noah Berger, Kaylee Chen, Juliana Dees, and Nate Garlati

We created a model to predict "Cinderella teams" in March Madness.

Part 1

Once Upon a time...

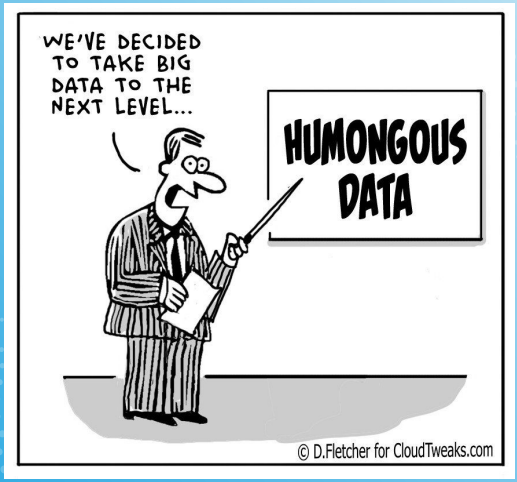
Stating our goal and Data Collection

Goal

We wanted to see if we could create a model to predict Cinderella teams in March Madness using regular season data.

Collecting Data

We collected our data from Sports Reference (regular season stats), Bracket Odds (predicted wins by seed), and Data World (March Madness results). We then combined these multiple data sets to create one new data set, ncaa_joint, so we could easily manipulate it to fit our needs.



Part 2

Defining the Glass Slipper

What makes a team a Cinderella?

A ‘cinderella team’ is defined by the NCAA as:

- A team that is the 11th seed or worse
- Wins at least 2 games in March Madness* (makes it to the sweet sixteen round)

*Does not include wild card games

THE DILEMMA

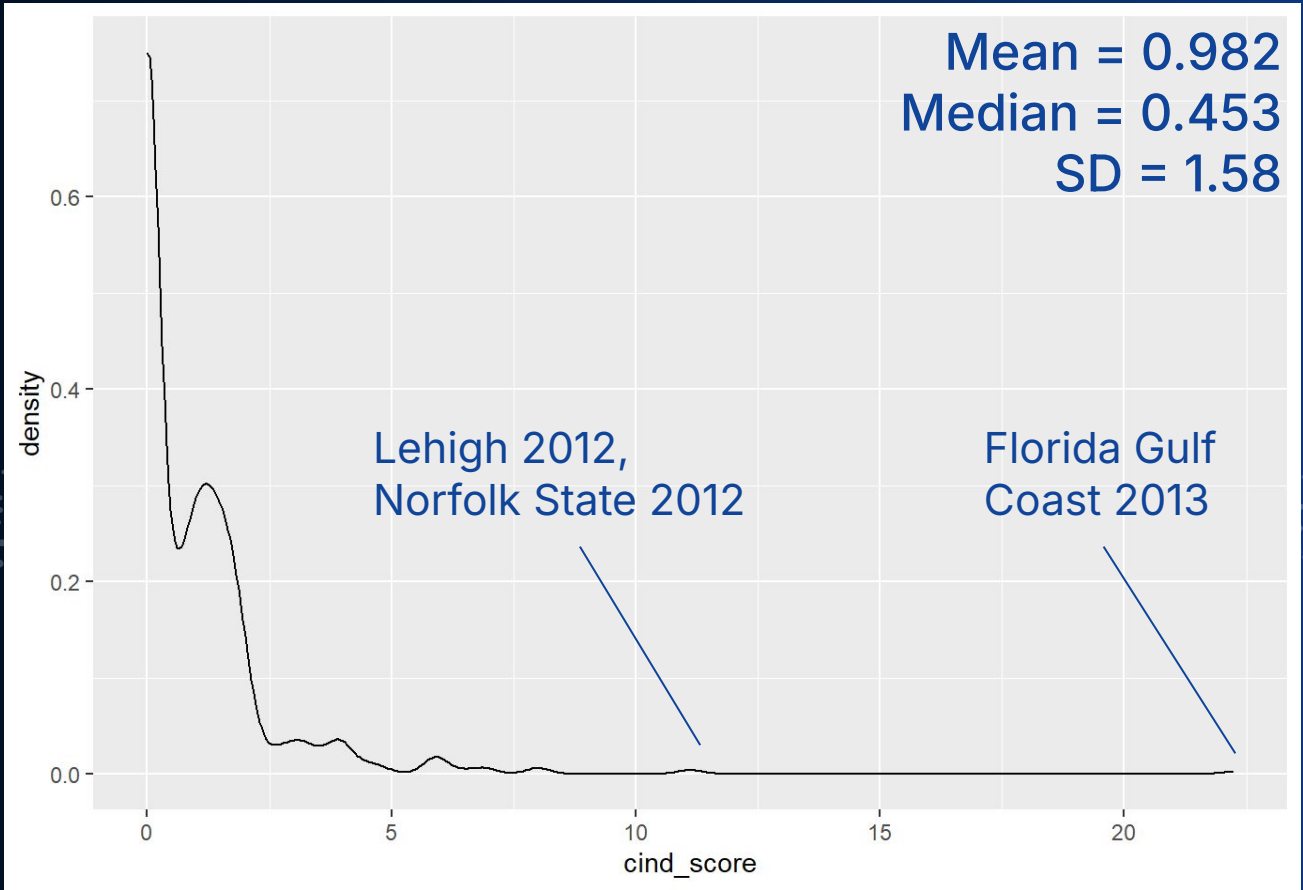
- Binomial = hard to model
- Continuous score
- Cinderella team = one that won more games than expected

THE SOLUTION

1. Find the predicted number of wins for each seed
2. Find the actual amount of games won by each team
3. Divide actual games won in March Madness by predicted number of wins
4. Call this new number the “Cinderella Score”
5. Final Formula: cind_score = mm_games_won/seed_predicted_wins

Seed	R64, R32, S16, E8, F4, NF	Expected Number of Wins
1	147, 126, 101, 60, 37, 24	3.34
2	138, 93, 67, 32, 13, 5	2.35
3	126, 77, 37, 17, 11, 4	1.84
4	117, 70, 22, 13, 3, 1	1.53
5	95, 50, 10, 7, 3, 0	1.11
6	91, 43, 15, 3, 2, 1	1.05
7	90, 28, 10, 3, 1, 1	0.90
8	72, 15, 9, 6, 4, 1	0.72
9	76, 7, 4, 1, 0, 0	0.59
10	58, 24, 9, 1, 0, 0	0.62
11	57, 26, 9, 5, 0, 0	0.66
12	53, 22, 2, 0, 0, 0	0.52
13	31, 6, 0, 0, 0, 0	0.25
14	22, 2, 0, 0, 0, 0	0.17
15	10, 3, 1, 0, 0, 0	0.09
16	1, 0, 0, 0, 0, 0	0.007

Density Curve of Cinderella Scores



Part 3

Gifts from our fairy godmother

This is a metaphor where our fairy godmother represents R and the gifts are analogous to variables that correlate to cinderella teams.

Creating a new non-pythag formula

For NCAA Basketball

pythag_general = (points_for-points_against)/(points_for+points_against)

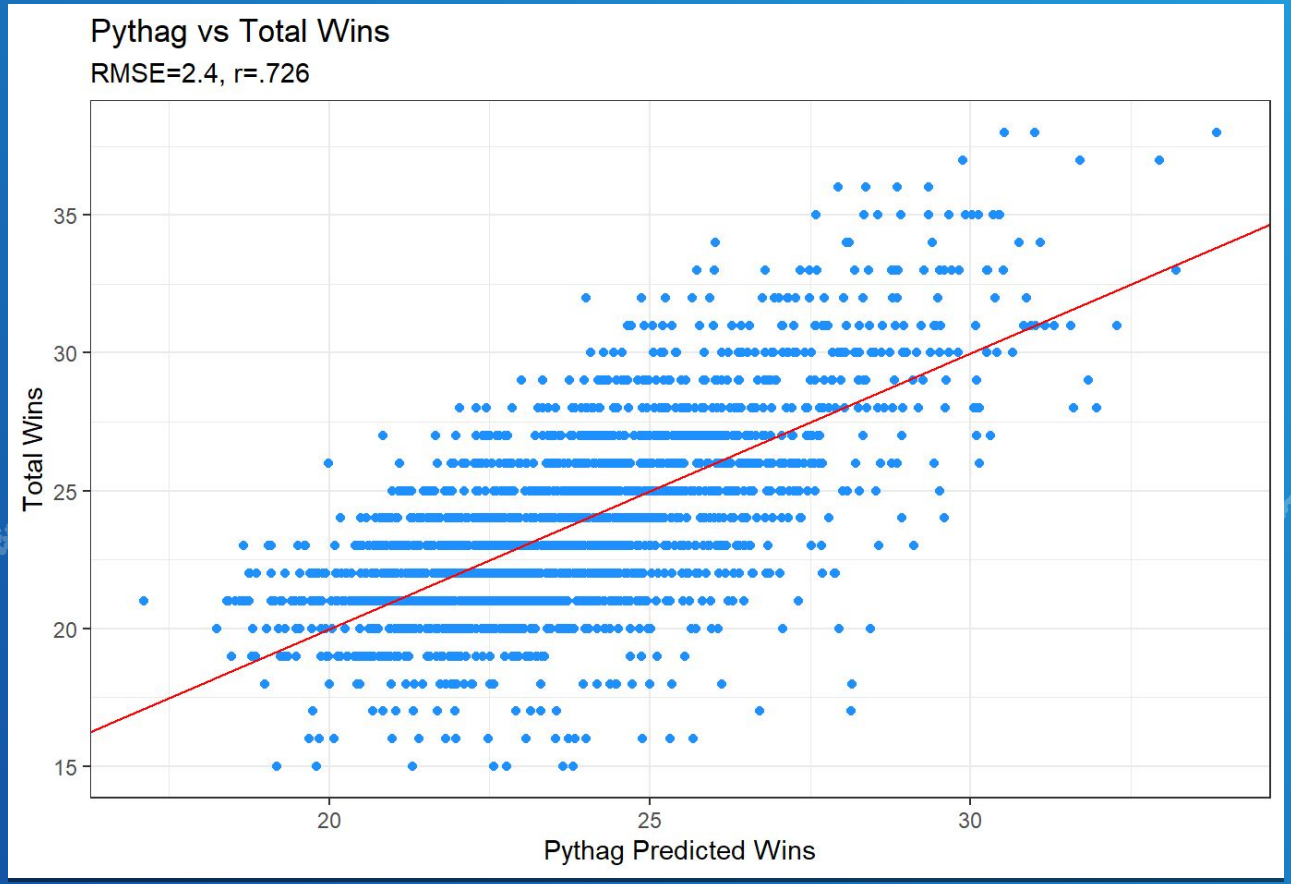
General formula

pythag_predicted_wins = 18.68 + 97.28(pythag_general)

New non-pythag formula for NCAA basketball

Pythag_resid = (pythag_predicted_wins)-(total_wins)

Formula for pythag residual



A chart representing the correlation of a team's expected wins using the Pythag formula, and their actual win total.

Other useful Variables

- SOS = Strength of Schedule
- SRS = Simple Rating System
- TPP = Three Point Percentage
- FGP = Field Goal Percentage

Part 4

The Clock Strikes 12

Our final model

Our Final Model

Phat_cind_score =

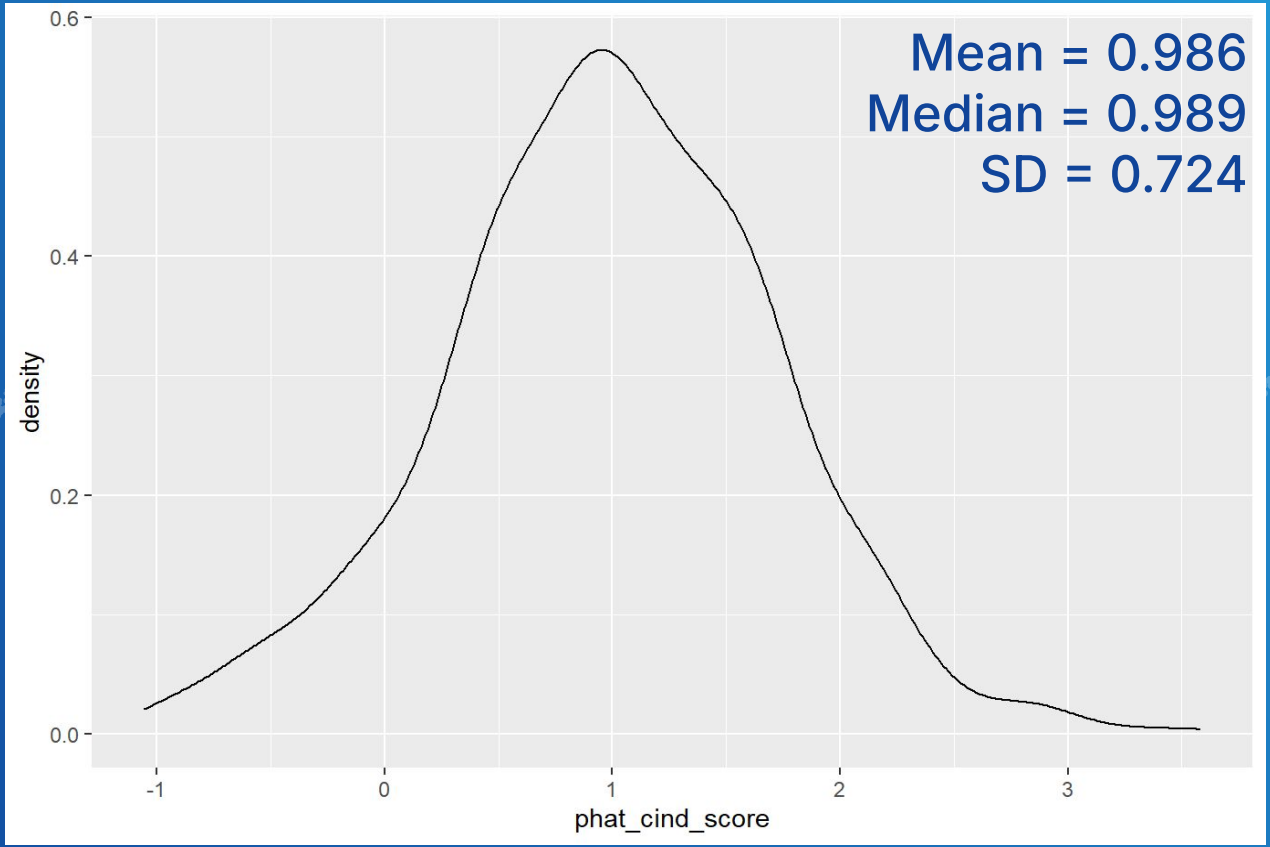
0.686543 + pythag_predicted_wins*0.238709 +

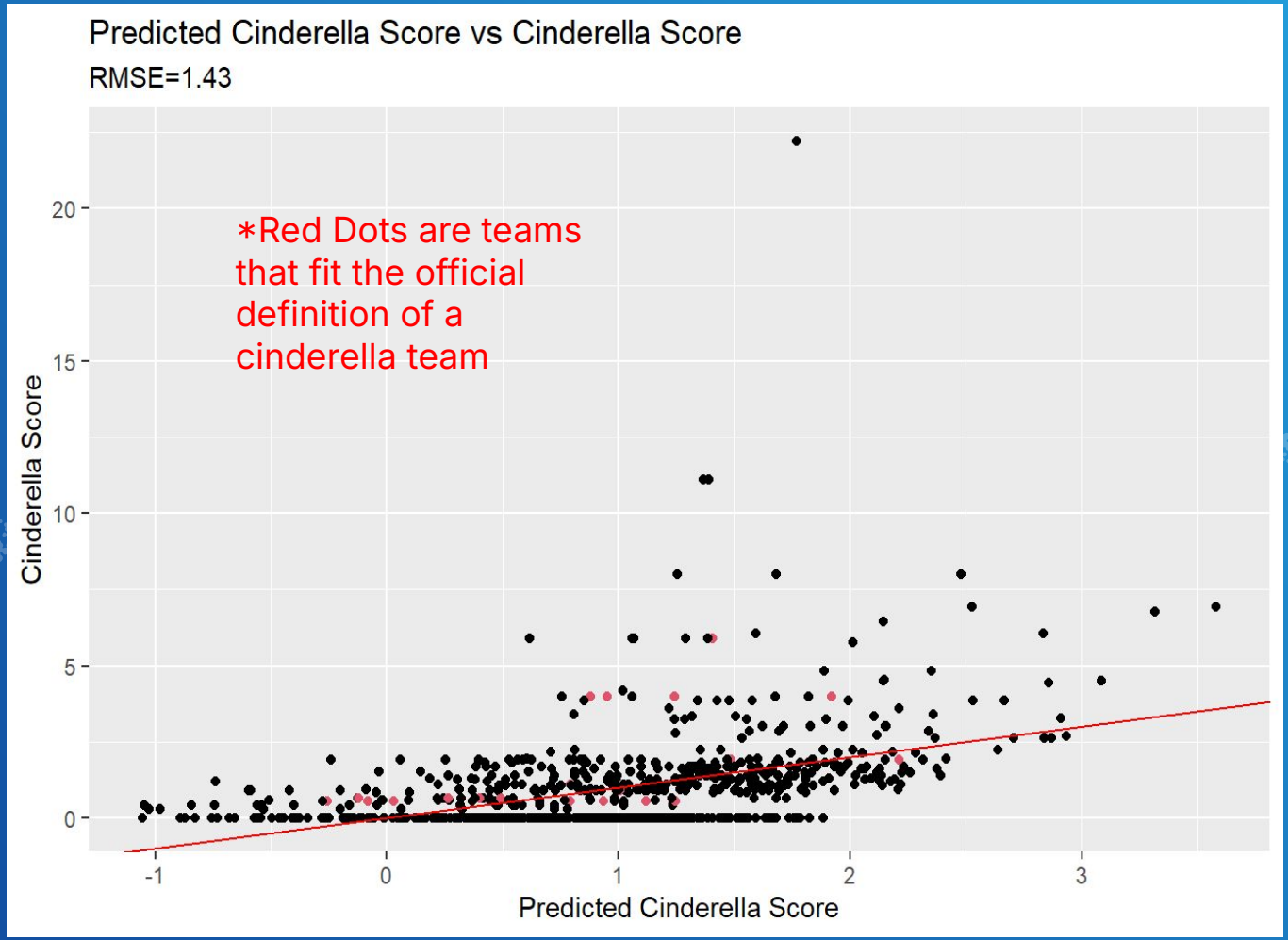
pythag_resid*0.393451 + WLP*-7.422530 +

seed_predicted_wins*-1.025192 + SOS*0.00863 +

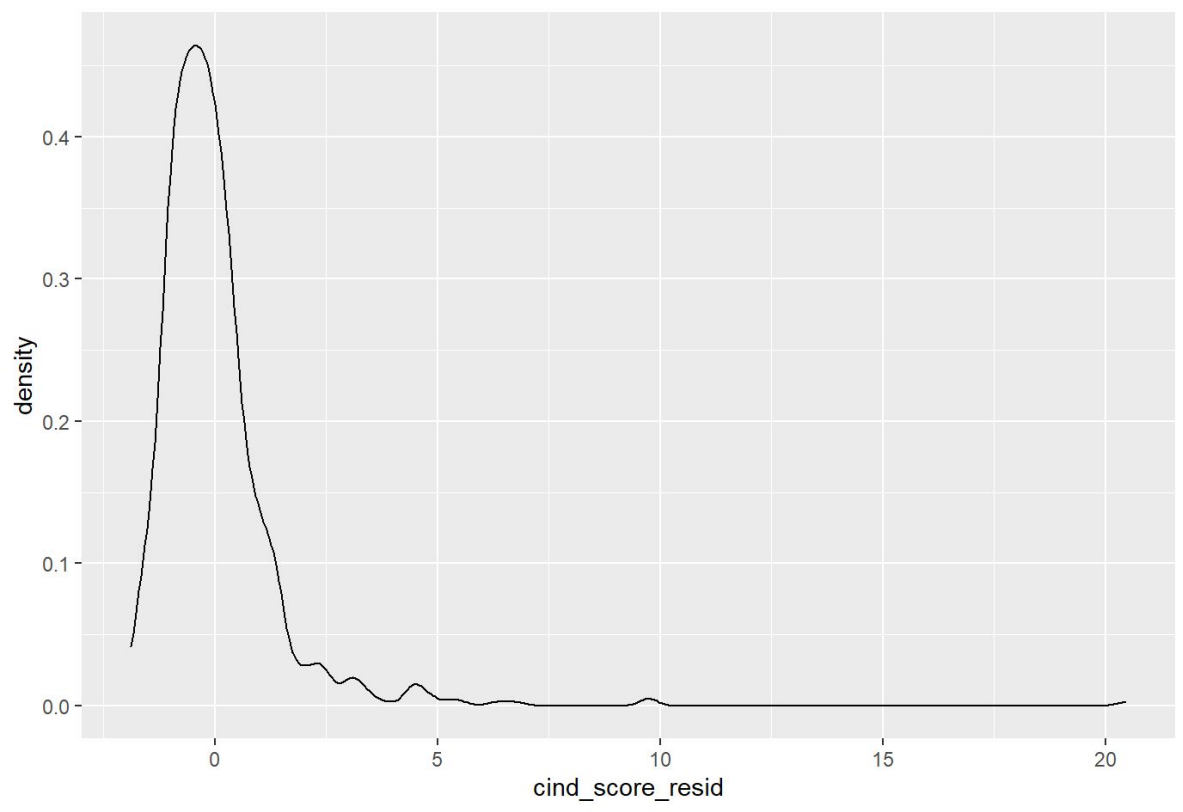
SRS*0.102137 + TPP*0.951688 + FGP*-1.94901

Density Curve of Predicted Cinderella Scores




Predicted Cinderella Score vs Cinderella Score
RMSE=1.43
*Red Dots are teams that fit the official definition of a cinderella team
Cinderella Score
20
15
10
5
0
-1
0
1
2
3
Predicted Cinderella Score

Density Curve of (Predicted Cinderella Score) - (Actual Cinderella Score)



Part 5

Happily Ever After

Final Conclusions

A couple of outliers

- St. Peter's 2022 has a predicted cinderella score of .837
- They had an *actual* cinderella score of 33.33
- UMBC 2018 has a predicted cinderella score of 1.25
- They had an *actual* cinderella score of 142.86

Conclusions

- extreme outliers, particularly in lower seeded teams
- extremely random and unpredictable
- Exponential formula?
- Rare events will take place over many trials

References and Resources

- Data from Sports Reference, Bracket Odds, and Data World
- Referenced library of past moneyball projects and moneyball project template
- Coding help from Stackoverflow and StatisticsGlobe
- Pictures from saintpeterspeacocks.com, Wikipedia, Hulu, and eonline.com
- Special thanks to Professor Wyner, Naomi, Justin, Elan, Jake and all the other TAs!

Thank you!

Questions?