# NCAA Men's Basketball Performance Analysis (2013–2025)

### **Course: Data Analysis & Visualization**

**Authors:** Keyu Chen, Deniz U Zeran  
**Date:** May 8, 2025

### **Executive Summary**

This project explores NCAA Division I Men's Basketball team performance between 2013 and 2025. By combining Tableau dashboards and Python-based clustering analysis, we aimed to identify performance patterns, highlight statistical trends, and investigate the phenomenon of "Cinderella teams" — low-seeded teams that make unexpectedly deep tournament runs. Tableau was used for interactive visual summaries, while Python was used to apply machine learning clustering techniques and data-driven Cinderella identification.

### **Table of Contents**

1. Introduction
2. Data Overview
3. Methodology
4. Exploratory Visualizations
5. Explanatory Data Analysis
6. Interactive Dashboard
7. Conclusions
8. References

### **1. Introduction**

* **Background:** NCAA March Madness is one of the most popular and unpredictable sports tournaments. Every year, teams defy expectations by upsetting higher seeds and making historic tournament runs. These teams are often referred to as "Cinderella teams."
* **Objective:** Our primary goal was to analyze the historical performance data of NCAA Men's Basketball teams to understand which statistical factors correlate with deep tournament success and to develop tools for identifying potential Cinderella teams.
* **Scope:** This study covers 12 years of NCAA team data (2013–2025), using both visual exploration and unsupervised machine learning techniques.

### **2. Data Overview**

* **Source**: Kaggle - College Basketball Dataset by Andrew Sundberg.

**Numeric Fields (Integer or Float):**

**Game Records:**

* G: Total number of games played
* W: Total number of wins

**Efficiency Metrics:**

* ADJOE: Adjusted offensive efficiency (points scored per 100 possessions)
* ADJDE: Adjusted defensive efficiency (points allowed per 100 possessions)
* BARTHAG: Power rating (probability of beating an average D-I team)

**Shooting Metrics:**

* EFG\_O: Effective field goal percentage (offense)
* EFG\_D: Effective field goal percentage allowed (defense)
* 2P\_O: Two-point field goal percentage (offense)
* 2P\_D: Two-point field goal percentage allowed
* 3P\_O: Three-point field goal percentage (offense)
* 3P\_D: Three-point field goal percentage allowed

**Turnovers, Rebounding & Free Throws:**

* TOR: Turnover rate (offense)
* TORD: Turnover rate forced (defense)
* ORB: Offensive rebound rate
* DRB: Defensive rebound rate
* FTR: Free throw rate (offense)
* FTRD: Free throw rate allowed (defense)

**Pace & Tournament Value:**

* ADJ\_T: Adjusted tempo (possessions per 40 minutes)
* WAB: Wins Above Bubble (relative to the NCAA tournament bubble)

**Tournament Information:**

* SEED: Seed number in the NCAA tournament
* YEAR: Season year (e.g., 2016, 2019)

### **3. Methodology**

**Tableau Analysis**  
 Five interactive visual dashboards were created:

1. Box Plot – BARTHAG by Seed
2. Scatter Plot – ADJOE vs. ADJDE
3. Heatmap – Seed vs. Postseason Result
4. Parallel Coordinates – Team Metric Comparison
5. Bar Chart – Tier vs. Postseason Outcome

### **Python Clustering Analysis**

* Selected 17 performance metrics
* Standardized using StandardScaler
* Applied KMeans clustering (k=5)
* Visualized using radar charts and PCA
* Applied Cinderella filter:  
  + SEED ≥ 10
  + POSTSEASON in [Elite 8, Final Four, Champion]

### **4. Exploratory Data Analysis**

Initial analysis revealed strong correlation between BARTHAG, ADJOE, and ADJDE with postseason success. Rebounding (ORB, DRB) and turnover metrics (TOR, TORD) also showed distinct patterns across seed tiers and conferences. The adjusted tempo (ADJ\_T) trend indicated a shift toward slower, more strategic play in recent years.

Correlation analysis showed EFG\_O and EFG\_D were key indicators of deep runs, underscoring the importance of efficient shot-making and shot defense.

#### **Clustering Analysis & Cinderella Detection**

To explore hidden performance patterns, we applied KMeans clustering to group postseason teams into five clusters based on 17 statistical features. Each cluster reflected unique strengths and weaknesses:

### **Cluster 0**

Typically around seed 9 with moderate to high WAB. Known for strong defensive disruption (TORD > 20%) and solid defensive rebounding—this is the core Cinderella zone.

### **Cluster 1**

Top seeds (1–3) with the highest WAB. Elite offensive and defensive efficiency (ADJOE and ADJDE), representing championship contenders.

### **Cluster 2**

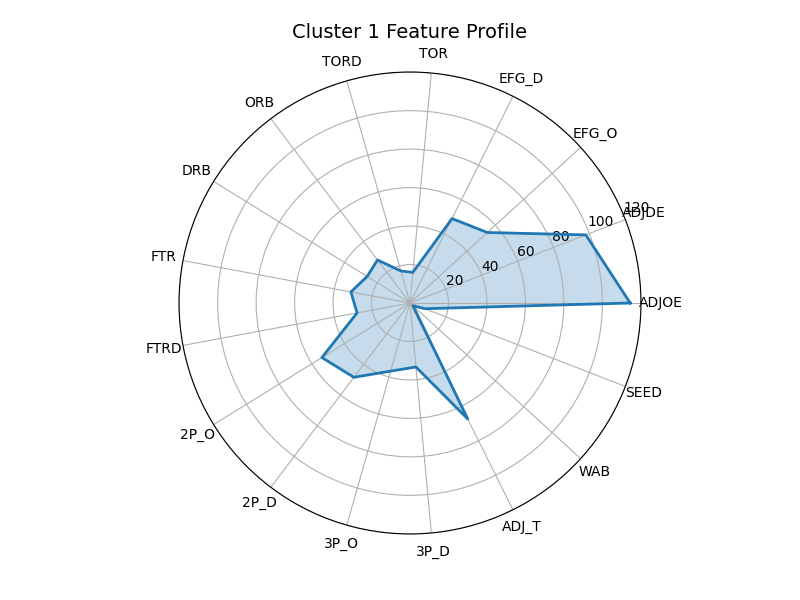
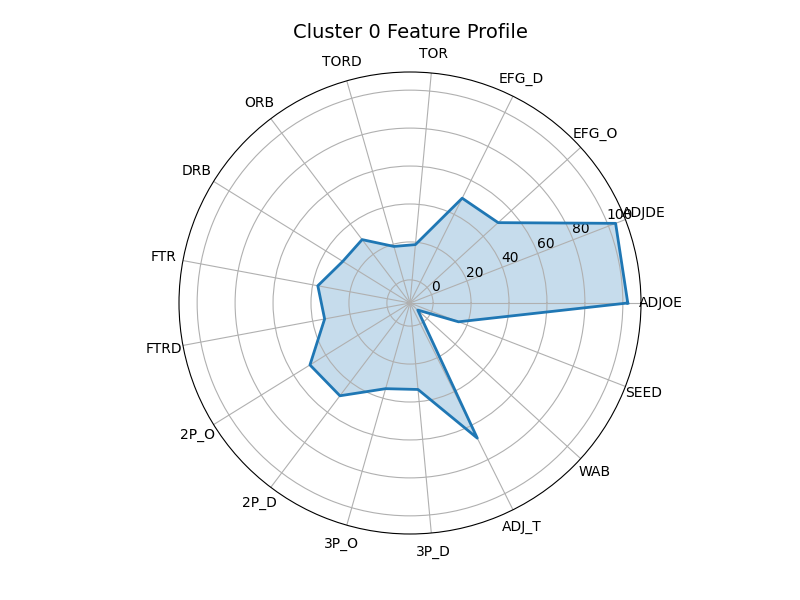
Mostly seeds above 15 with the lowest WAB. Weak across all metrics, poor defense—frequent early exits.

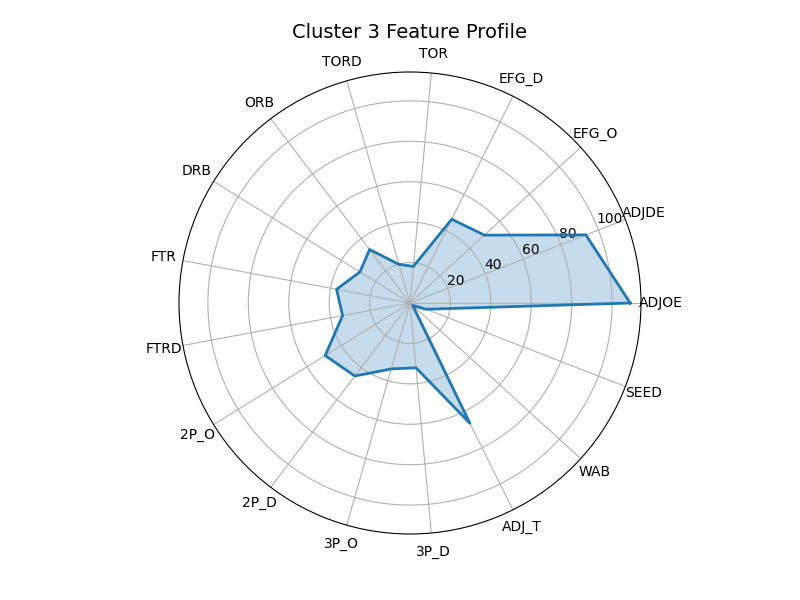
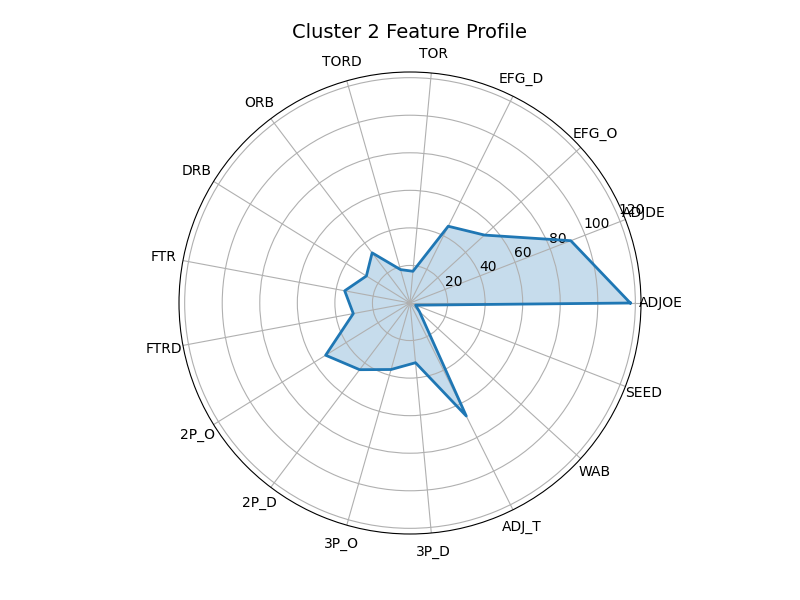
### **Cluster 3**

Mid-tier seeds (4–9) with moderate WAB. Balanced but unremarkable stats, representing stable but unspectacular teams.

### **Cluster 4**

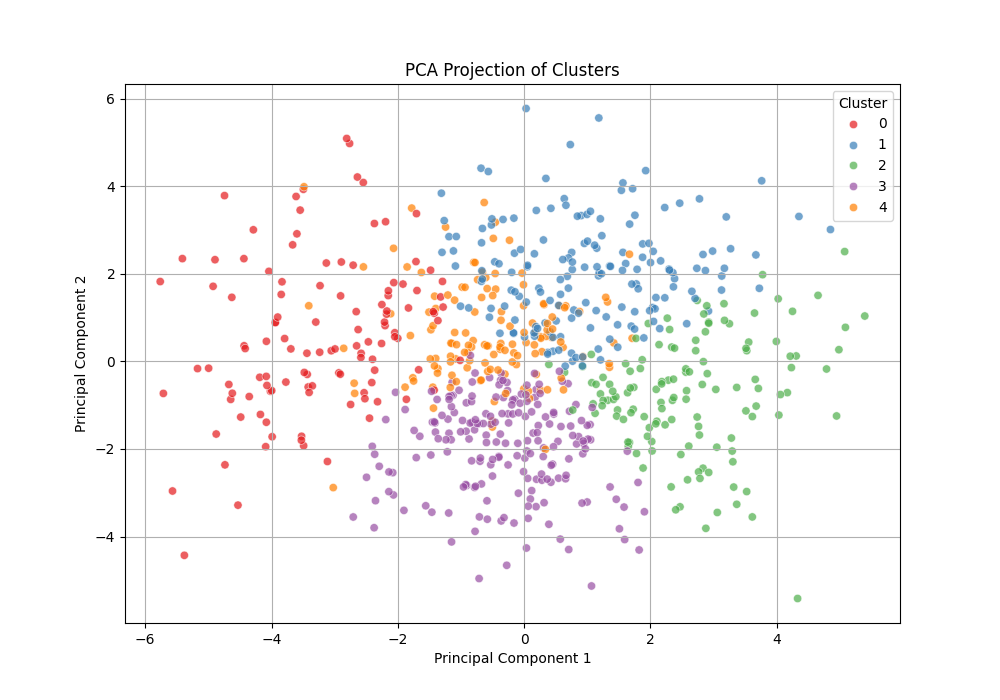
Roughly around seed 11 with near-zero WAB. Similar to Cluster 0 but with less consistent overperformance—some Cinderella cases, mixed outcomes.







#### **PCA Cluster Projection**



We used Principal Component Analysis (PCA) to reduce the 17 statistical metrics into two dimensions and visualize the clustering structure.

* **Cluster 1** (blue) appears tightly grouped in the upper-right quadrant — representing high-performing, consistent teams with elite metrics.
* **Cluster 2** (green) is more dispersed across the lower-right area — these teams generally have weak statistics and tend to exit the tournament early.
* **Cluster 3** (purple) occupies the central space — reflecting balanced but statistically average teams without distinctive strengths.
* **Cluster 0** (red) is positioned on the far left — indicating teams that generate strong defensive disruption, force turnovers, and often exceed expectations.
* **Cluster 4** (orange) is located near the center — it shares several traits with Cluster 0 but lacks the same consistency, making it a mixed group with potential Cinderella outcomes.

This PCA visualization reinforces the results of our radar chart analysis. It shows that Cinderella teams are not random anomalies but belong to identifiable statistical profiles characterized by defensive intensity, rebound strength, and above-bubble performance.

### **Cinderella Traits**

An analysis of teams from Clusters 0 and 4 that achieved deep postseason success reveals a consistent set of distinguishing performance metrics:

* **Wins Above Bubble (WAB):** A measure of how significantly a team outperformed expectations relative to its tournament seeding. High WAB values are a common characteristic of successful underdog teams.
* **Turnover Rate Forced (TORD):** Reflects a team’s ability to create defensive pressure and disrupt opponent possessions—an essential trait of defensively oriented Cinderella teams.
* **Defensive Rebound Rate (DRB):** Indicates a team’s effectiveness at preventing second-chance scoring opportunities, reinforcing the importance of defensive control.
* **BARTHAG:** A composite power rating that captures a team’s overall quality and consistency, estimating its probability of defeating an average Division I opponent.

These metrics collectively suggest that successful Cinderella teams are defined less by offensive firepower and more by their defensive discipline, tactical efficiency, and ability to exceed expectations through structured and resilient play. Their performance is not anomalous but reflects a measurable and repeatable pattern.

#### **Identified Cinderella Teams**

Using the following criteria:

* Belonging to Cluster 0 or 4
* Seed ≥ 10
* Postseason result of E8, F4, or Champions

We identified five classic Cinderella runs:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Team | Year | Seed | Postseason | Cluster | BARTHAG |
| Xavier | 2017 | 11 | E8 | 4 | 0.8713 |
| Dayton | 2014 | 11 | E8 | 0 | 0.8188 |
| Syracuse | 2016 | 10 | F4 | 0 | 0.8857 |
| Loyola Chicago | 2018 | 11 | F4 | 4 | 0.8444 |
| UCLA | 2021 | 11 | F4 | 4 | 0.8350 |

The analysis identified five teams that met the established criteria for Cinderella classification: inclusion in Cluster 0 or 4, a tournament seed of 10 or higher, and advancement to at least the Elite Eight. Although seeded as underdogs, these teams consistently demonstrated notable statistical characteristics—specifically, strong defensive efficiency, elevated Wins Above Bubble (WAB) values, and BARTHAG ratings comparable to top-tier programs.

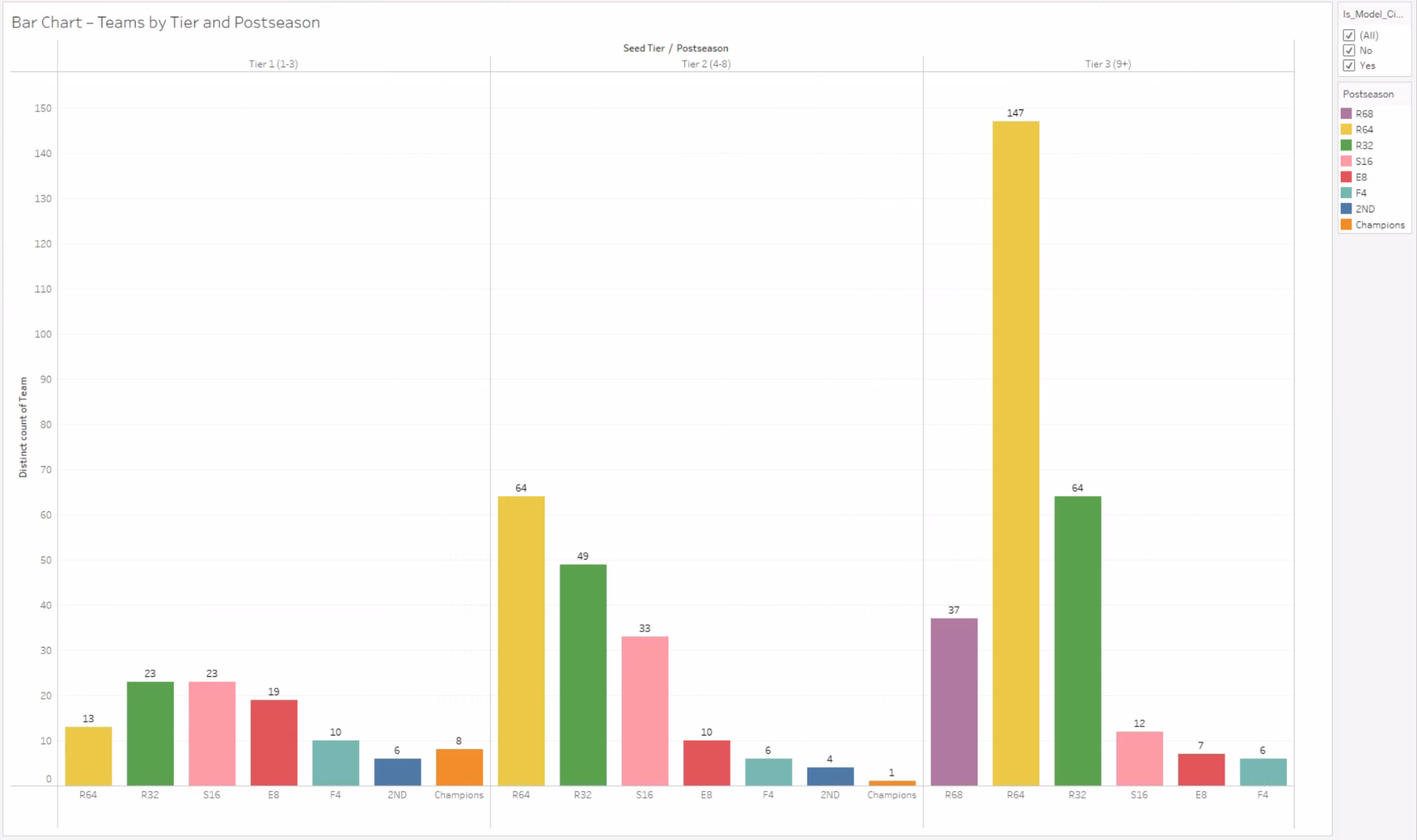
The recurrence of these traits across multiple seasons and conferences suggests that their success was not incidental. Rather, it reflects a statistically identifiable and repeatable performance profile. This consistency lends credibility to the clustering framework and supports its potential application beyond retrospective analysis.

The model thus offers a practical, data-informed approach for evaluating lower-seeded teams prior to tournament play. By identifying those that align with historically successful profiles, it becomes possible to anticipate which underdog teams are most likely to advance beyond expectations and generate significant upsets.

## **5. Explanatory Visualizations (Tableau Analysis)**

To determine what makes a Cinderella team succeed, we created a series of Tableau visualizations that analyze postseason performance, statistical traits, and team behavior across the 2013–2025 NCAA Men’s Basketball Tournaments. Each visualization contributes a piece of the answer to our guiding question: **what metrics truly define a Cinderella team's path to success?**

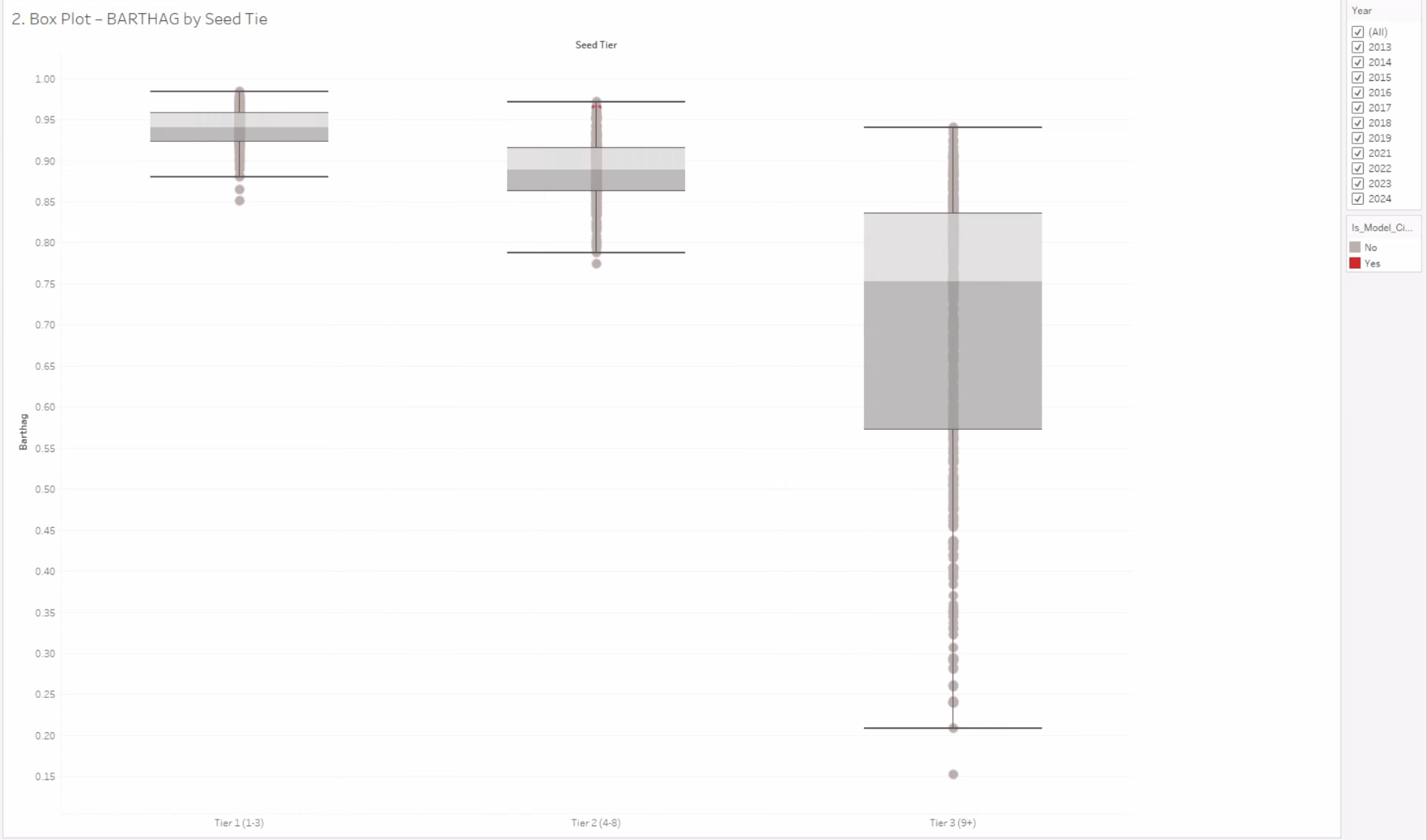
### **Bar Chart – Tier vs. Postseason Performance**



**Purpose:** This chart counts how many teams from each seed tier (Tier 1: Seeds 1–3, Tier 2: Seeds 4–8, Tier 3: Seeds 9+) advanced to each round of the tournament.

**Insight:** Tier 3 teams—our primary Cinderella candidates—rarely win championships, but frequently reach the Sweet 16, Elite Eight, and even Final Four. Tier 2 teams occasionally win the tournament, suggesting that deep runs are not exclusive to top seeds. This justifies focusing our analysis on Tier 2 and Tier 3 as potential Cinderella categories and validates our Cinderella filter criteria.

### **Box Plot – BARTHAG by Seed Tier**



**Purpose:** We used this box-and-whisker plot to compare the distribution of BARTHAG (a power rating estimating a team's chance of beating an average opponent) across seed tiers.

**Insight:** While Tier 1 teams predictably have the highest median BARTHAG, some Tier 2 and Tier 3 teams overlap significantly with Tier 1. This implies that many Cinderella teams are **statistically stronger than their seed suggests**. For builders and analysts, it indicates that a low BARTHAG is a red flag—but a mid-to-high BARTHAG (> 0.83) is a potential Cinderella trait, even in Tier 3.

### **Scatter Plot – ADJOE vs. ADJDE**



**Purpose:** We visualized each team’s offensive (ADJOE) vs. defensive (ADJDE) efficiency to understand their playing style and compare Cinderellas to Champions.

**Insight:** Champions cluster in the top-right quadrant—elite on both offense and defense. Cinderella teams, however, tend to lean toward **strong defense (low ADJDE)** with **average-to-above-average offense**. This reinforces the idea that Cinderella teams win not by overpowering but by **controlling tempo, disrupting rhythm, and limiting scoring**. The key takeaway: **building a Cinderella team starts with an elite defense, not a flashy offense**.

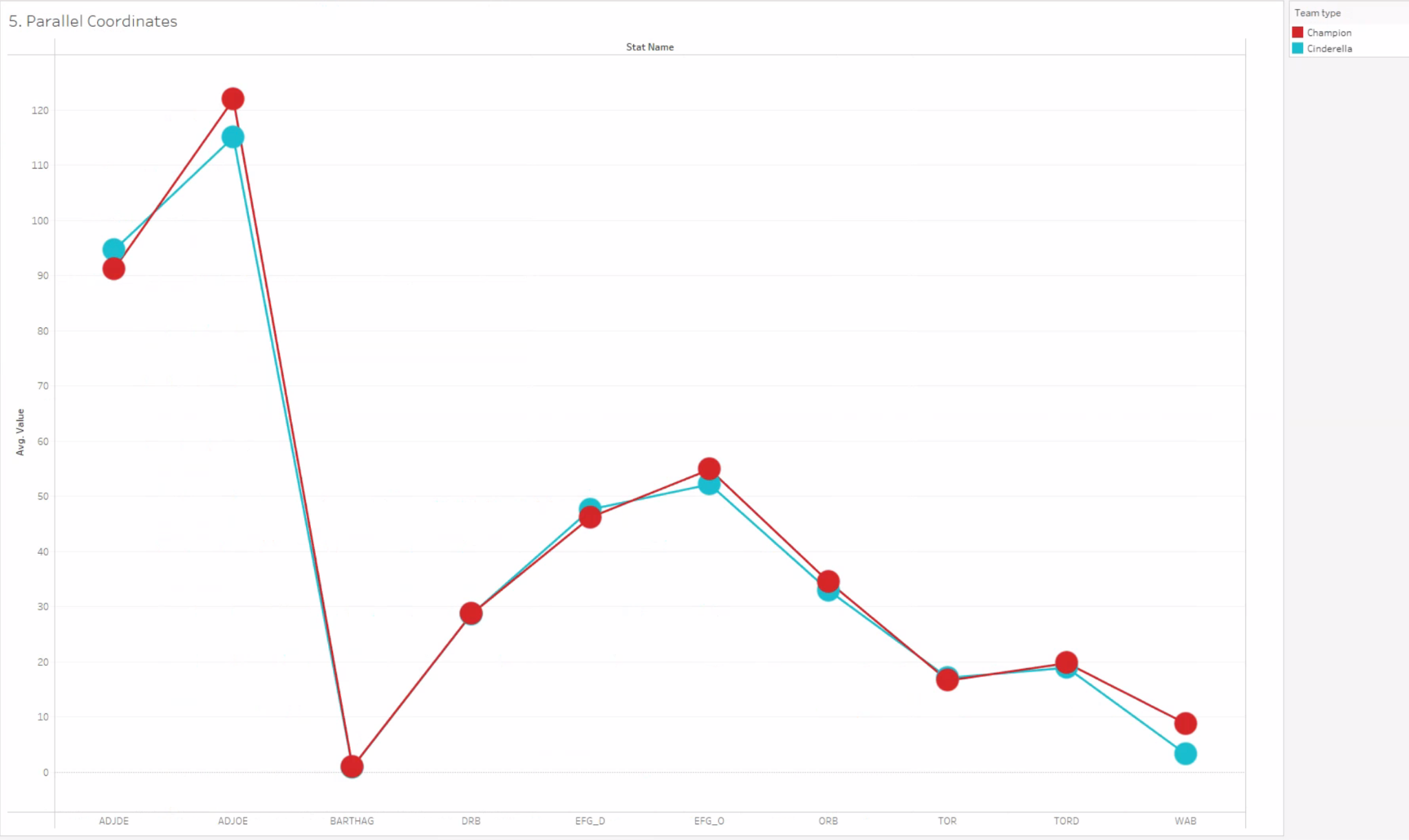
### **Heatmap – Seed vs. Postseason Round**



**Purpose:** This matrix shows which seeds historically reach which postseason rounds, with cell color intensity representing frequency.

**Insight:** Seeds 10, 11, and 12 have the most Cinderella activity—making it as far as the Elite Eight and Final Four on multiple occasions. These seeds combine **underrated strength** and **favorable matchups**, suggesting that scouting lower seeds in this range—especially those with strong advanced metrics—is critical for Cinderella identification. This view also highlights how rare it is for seeds above 13 to break through, suggesting a practical cutoff for most models.

### **Parallel Coordinates – Multi-Metric Comparison**



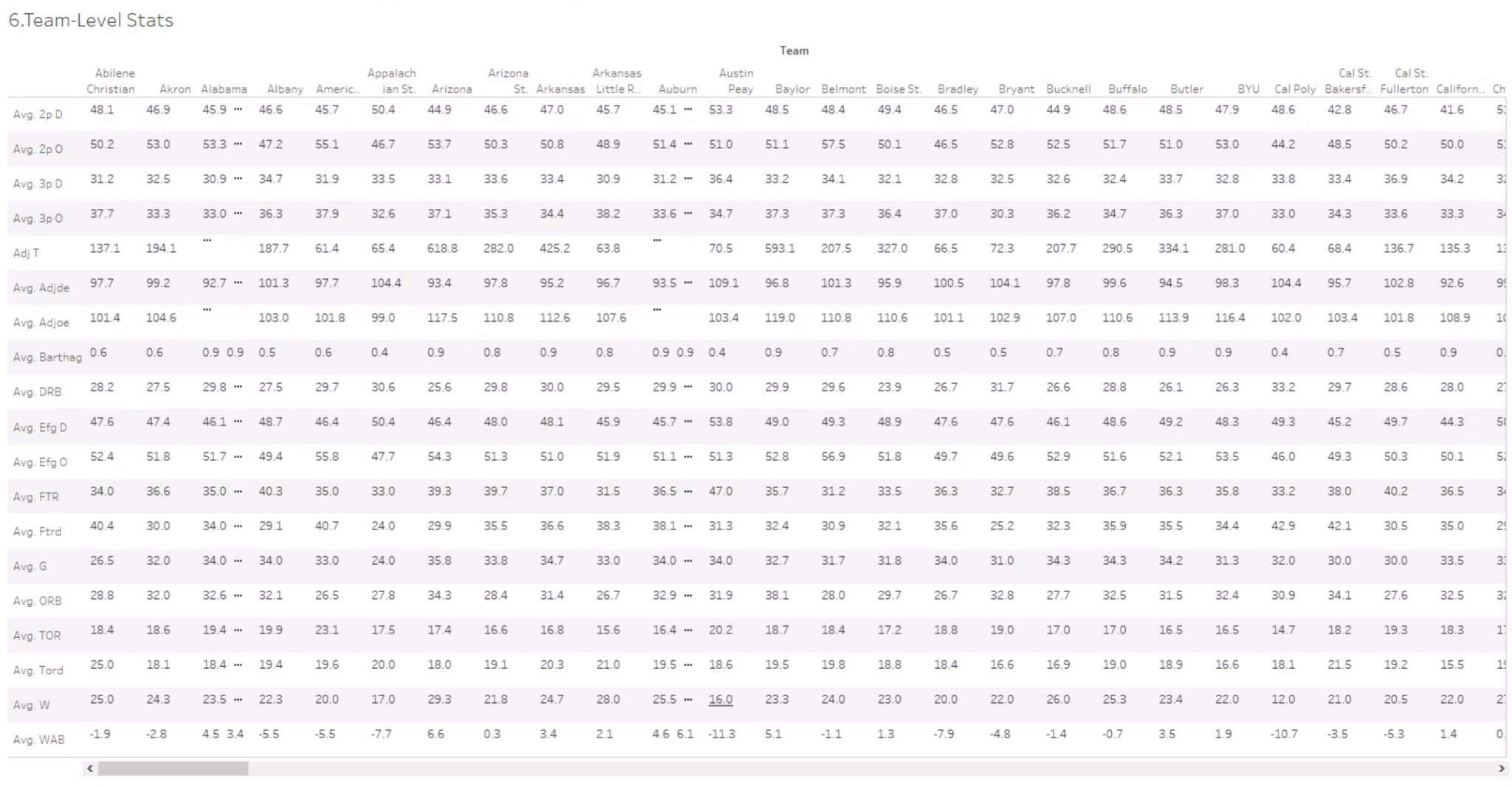
**Purpose:** This chart compares average values across critical metrics—BARTHAG, WAB (Wins Above Bubble), TORD (Turnover Rate Forced), and DRB (Defensive Rebound Rate)—for Cinderellas, Champions, and Other teams.

**Insight:** Cinderellas consistently show:

* **High WAB** (2.5–4.0): They outperformed expectations.
* **High TORD**: Cinderella teams **generate chaos**—they force more turnovers than both Champions and Other teams.
* **Strong DRB**: They **control the glass defensively**, limiting second-chance points.
* **Moderate-to-high BARTHAG**: They are far better than their seed implies.

This chart crystalizes the Cinderella formula:

**Take a high-WAB team that plays tough defense (TORD, DRB) and has a BARTHAG north of 0.83—and you have the statistical skeleton of a Cinderella.**



### **Team-Level Stats Table**

**Purpose:** This interactive table provides a team-by-team breakdown of all statistical metrics, updated based on filters (e.g., seed, postseason, team type).

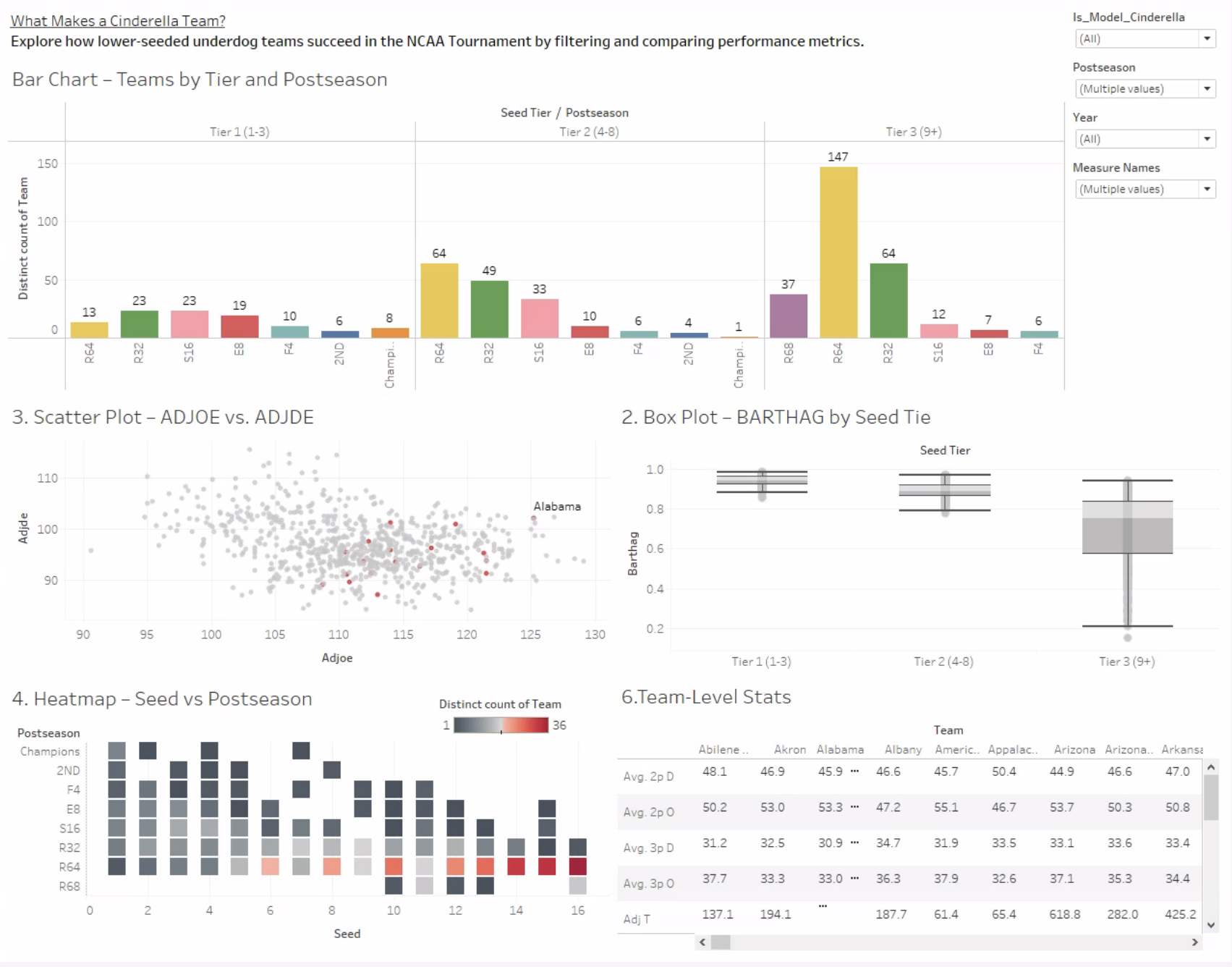
**Insight:** It lets users validate the macro-level patterns by examining individual Cinderellas (e.g., 2018 Loyola Chicago, 2021 UCLA). The table confirms that many of these teams exhibit **the exact statistical traits surfaced in the charts above**, giving credibility to the model and helping identify future candidates.

### **Summary Recommendation**

Based on these Tableau visualizations, we recommend focusing on the following metrics when evaluating or building a potential Cinderella team:

|  |  |
| --- | --- |
| **Metric** | **Why It Matters** |
| **BARTHAG > 0.83** | Indicates statistical parity with top teams |
| **WAB > 2.5** | Suggests team significantly outperformed expectation |
| **TORD > 20%** | Points to defensive disruptiveness |
| **DRB > 70%** | Critical for limiting opponent scoring chances |
| **Low ADJDE** | Defensive identity is the common Cinderella edge |

### **6. Interactive Dashboard**

  
  
To synthesize our visual findings and support deeper exploration, we developed a fully interactive Tableau dashboard titled **“What Makes a Cinderella Team?”** The dashboard integrates key statistical visualizations and user-driven filtering to help identify the traits that characterize successful NCAA underdog teams. Rather than displaying static insights, it functions as a dynamic tool for analyzing tournament performance through multiple lenses.

#### **Dashboard Structure and Components**

The layout progresses from high-level patterns to individual team insights. Each element updates based on user interaction, allowing real-time filtering and hypothesis testing. The dashboard includes the following key components:

* **Bar Chart – Seed Tier vs. Postseason Round:** This chart functions as the central interactive filter. Users can select a specific seed tier and postseason round to dynamically update all other charts and the team table. It visually confirms that Tier 2 and Tier 3 teams—often classified as Cinderella candidates—frequently reach the Elite Eight, Final Four, or even the Championship game.
* **Box Plot – BARTHAG by Seed Tier:** This visualization shows the distribution of team power ratings across seed groups. It highlights that lower-seeded teams occasionally match Tier 1 teams in BARTHAG, suggesting that seeding may not always reflect true performance.
* **Scatter Plot – ADJOE vs. ADJDE:** This plot compares offensive and defensive efficiency. Cinderella teams tend to cluster in the quadrant of strong defense (low ADJDE) and average offense, whereas champions generally show strength in both dimensions.
* **Heatmap – Seed vs. Postseason Round:** This matrix reveals which specific seeds most frequently reach deeper rounds. Seeds 10, 11, and 12 emerge as consistently volatile, further supporting our definition of Cinderella seeds.
* **Team-Level Stats Table:** The table displays detailed metrics for each team and updates in real time based on selected filters. It enables users to examine whether specific teams meet the Cinderella statistical profile.

#### **Interactive Features**

The dashboard is designed to facilitate intuitive, layered analysis through:

* **Central Filtering via Bar Chart:** Users can click any combination of seed tier and tournament round to filter all other views. This makes it possible to isolate patterns such as “Final Four teams seeded Tier 3” and examine their common characteristics.
* **Global Filters:** Additional filters include Year, Team, Seed Tier, Postseason Round, and Cinderella Status. These allow granular control over the dataset and ensure consistent filtering across all visualizations.
* **Real-Time Validation:** Users can move from summary visuals to the team-level stats table to verify whether outliers or exceptions follow the identified patterns.

#### **Purpose and Value**

This dashboard bridges the gap between exploratory visualizations and practical analysis. It empowers users to investigate the hypothesis that Cinderella teams share measurable statistical traits—namely, high WAB, strong defensive efficiency (TORD), and effective rebounding (DRB). The dashboard also enables:

* Real-time comparisons between Cinderella teams and champions
* Year-over-year pattern recognition
* Team-specific deep dives to support or challenge bracket decisions

In short, the dashboard is not merely a visual summary—it is a decision-support tool grounded in repeatable, metric-driven logic.

## **7. Conclusions**

This project applied two complementary approaches to analyze NCAA tournament performance and uncover potential Cinderella teams:

* **Tableau visual analysis** (by Deniz), offering interactive insights and intuitive pattern recognition.
* **Python-based clustering and modeling** (by Kevin), providing structured, data-driven segmentation.

Together, these methods offered both breadth and depth in evaluating underdog tournament runs.

### **Overlapping Findings**

Several well-known Cinderella teams appeared across both methods, reinforcing their validity:

* Loyola Chicago (2018) – 11-seed, Final Four
* UCLA (2021) – 11-seed, Final Four
* Syracuse (2016) – 10-seed, Final Four

These teams consistently exhibited traits such as strong defensive efficiency, solid rebounding, and high Wins Above Bubble (WAB) relative to their seeding.

### **Method Divergences: The UConn 2014 Case**

One key divergence was UConn’s 2014 championship run as a 7-seed:

* Appeared in Tableau scatterplots and heatmaps due to its tournament outcome.
* Was excluded from Python clustering results because the original filter required SEED ≥ 10.

This highlights a trade-off: Tableau enables visual storytelling and exception spotting, while Python provides consistent, criteria-driven results that may miss outliers without threshold adjustments.

### **Additional Discrepancies: Tableau Misses Dayton & Xavier**

Conversely, Python detected some statistically valid Cinderella teams not highlighted in Tableau, including:

* Dayton (2014) – 11-seed, Elite Eight
* Xavier (2017) – 11-seed, Elite Eight

This shows that Tableau’s effectiveness depends on how data is filtered or labeled. In contrast, Python clustering evaluated the full dataset and surfaced patterns based purely on statistical similarity.

### **Strengths & Limitations**

|  |  |  |
| --- | --- | --- |
| **Method** | **Strengths** | **Limitations** |
| Tableau | Intuitive and exploratory; good for spotting patterns and exceptions | May omit teams not highlighted manually or excluded by filters |
| Python (ML) | Comprehensive and consistent; ideal for statistical segmentation | Requires careful filter design; may exclude notable exceptions |

### **Key Traits of Successful Cinderella Teams**

Across both methods, a consistent statistical profile emerged. The following metrics define teams that most often exceed expectations:

* **WAB > 2.5** – Outperforms expectations relative to tournament bubble.
* **TORD > 20%** – Applies pressure and generates extra possessions.
* **DRB > 70%** – Limits opponent second-chance scoring.
* **BARTHAG > 0.83** – Indicates overall statistical strength despite lower seed.

These traits were most common among Cinderellas identified by both Tableau and Python clustering. They offer a practical foundation for evaluating or even predicting future Cinderella candidates.

**Successful Cinderella teams aren’t random. They are strong defenders, efficient rebounders, and high-performing teams overlooked by their seed.**

### **Final Takeaway**

The greatest value of this project lies in how the two methods complement each other:

* Tableau highlighted standout stories like UConn 2014, even when data filters would have excluded them.
* Python clustering surfaced overlooked but statistically credible Cinderellas like Dayton and Xavier.

Together, they offer a fuller picture of NCAA tournament dynamics—blending visual intuition with data-driven discovery. This dual-method strategy provides not only a retrospective analysis, but also a practical toolkit for identifying future tournament surprises.

### **8. References**

Sundberg, A. (2022). *College Basketball Dataset*. Kaggle.

Retrieved from: https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset