Unveiling Fan Bias in NCAA Bracket Predictions:

Trends, Accuracy, and the Impact of School Affinity

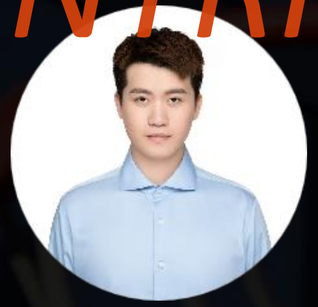
MARCH MADNESS

Team: **Win and Retired**

Huangkun Chen
Yingtong Wang
Nikhil Ram Atluri
Yingjie Chen

2025.2

TEAM ROLES & CONTRIBUTIONS



Huangkun Chen

Leader
Presentation
Tableau Analysis
Kaggle Submission

Email

chen5180@purdue.edu

Yingtong Wang

Presentation
Tableau Analysis
Kaggle Submission
Task Allocation

Email

wang6679@purdue.edu

Yingjie Chen

Tableau Analysis
Communication

Email

chen5301@purdue.edu

Nikhil Ram Atlurin

Tableau Analysis
Kaggle Submission

Email

atlurin@purdue.edu

We are all Purdue Univeristy MSBAIMer !

CCAC

TABLE OF CONTENTS

CCAC

PART 1

Introduction & Problem Framing

Problem Statement

Key Assumptions

Success Metrics

CCAC

PROBLEM STATEMENT

“ Our goal is to develop a **predictive model** based on fan-submitted bracket data to forecast the winners of the NCAA semifinals and national championship. Beyond that, we aim to analyze whether **school affinity influences fan predictions**—specifically, whether fans are more likely to support certain teams based on **geographic proximity** or historical loyalty. ”

ADDITIONAL ANALYTICS

Do prediction trends change over time?
(Macro-level trend)
↓

Is there significant fan bias in predictions?
(Core factor driving prediction behavior)
↓

Do fan predictions accurately reflect a team's actual competitive strength?
(Impact of prediction bias)
↓

Does home vs. away performance further influence fan predictions?
(Additional external factor)

KEY ASSUMPTIONS

DATA ASSUMPTIONS

1. Fan-submitted brackets accurately represent **real prediction behavior**.
2. The selected features **sufficiently describe** fan prediction patterns.
3. Data distributions remain **consistent** between training and testing sets.
4. Each fan's prediction is **independent** of others.

MODELING ASSUMPTIONS

1. **XGBoost** are suitable for this classification task.
2. Feature engineering improves predictive performance.
3. The model is robust to overfitting.
4. **AUC-ROC** is appropriate evaluation metrics.

SUCCESS METRICS

1. Predictive Model Performance Metrics (Evaluating Model Accuracy)

✅ **Accuracy** – Measures the percentage of correct predictions. Higher accuracy indicates better model performance.
✅ **AUC-ROC** (Area Under the Curve - Receiver Operating Characteristic) – Measures the model's ability to differentiate between different outcomes. AUC closer to 1.0 means better classification.

2. Fan Behavior & Bias Quantification Metrics (Evaluating Insights on Fan Predictions)

✅ **Bias Score** – Quantifies fan bias toward certain teams (e.g., overestimating popular teams).
✅ **Prediction Deviation** – Compares fan predictions vs. actual game results, revealing overestimated/underestimated teams.
✅ Regional Influence – Measures how geographic factors (DMA, clustering) impact fan predictions.

PART 2

Data & Feature Engineering

Data Source

Data Preprocessing

Data Relationships

DATA SOURCES

Training Set (bracket_train.csv)
Test Set (bracket_test.csv).
Submission Template (submission_template.csv)
CCAC 2025 - Institutions.csv
A list of schools and their associated school IDs.
Can be used for feature engineering, such as team strength, historical performance, or conference affiliations.
CCAC 2025 - Data Dictionary.xlsx
Provides detailed descriptions of each column in the dataset.
Essential for understanding feature meanings and ensuring accurate data preprocessing.
Division I Women_s Basketball Contests.csv

Key Features

Feature Category	Key Feature	Purpose	Usage Stage (Predictive Model / Visualization Analysis)
Demographic & Regional	CustomerDMACode	Represents the DMA code where the fan is located	Predictive Model & Visualization
	CustomerDMADescription	The regional name corresponding to the DMA code	Visualization
	CustomerPostalCodeLatitude & CustomerPostalCodeLongitude	Fan's geographic coordinates for regional analysis	Predictive Model & Visualization
	DMA_SuccessRate	Historical success rate of teams in a given DMA	Predictive Model
	LocationCluster	Geographic fan segmentation using K-Means clustering	Predictive Model & Visualization
Bracket-Specific	SemifinalWinner_East_West	Fan-predicted winner of the East vs. West semifinal	Predictive Model (Target Variable)
	SemifinalWinner_South_Midwest	Fan-predicted winner of the South vs. Midwest semifinal	Predictive Model (Target Variable)
	NationalChampion	Fan-predicted NCAA national champion	Predictive Model (Target Variable)
Historical Performance	BracketEntryId	Unique identifier for each fan's prediction bracket	Visualization
	MostPopularTeam	Most frequently predicted national champion in each DMA	Visualization

DATA PREPROCESSING

Preprocessing Step	Action Taken	Purpose
Categorical Encoding	Converted categorical variables (DMA Code, Winners) using Label Encoding & **One-Hot** Encoding	Makes categorical features suitable for machine learning models
Missing Value Handling	Applied **mean** imputation for missing geolocation & DMA success rate	Prevents loss of important data
Feature Engineering	Added DMA success rate, **K-Means** clustering (LocationCluster), and Most Popular Team per DMA	Enhances predictive power by including regional trends
Dataset Standardization	**Removed** unnecessary columns, structured data for modeling & visualization	Ensures dataset is clean and model-ready

DATA RELATIONSHIPS

The **Training dataset** contains Region Winners, Semifinal Winners, and National Champion as **Institution IDs**, which can be linked to the **Institution dataset** (CCAC 2025 - Institutions.csv) to retrieve school names and additional attributes such as conference affiliation and historical performance. Additionally, the DMA information in the Training dataset can be matched with the **Customer DMA dataset**, providing insights into regional fan preferences. The Submission template ensures that predictions align with the required format, and the **Data Dictionary** helps in understanding feature definitions for accurate preprocessing and modeling.

CCAC

PART 3

Predictive Modeling & Analysis

Predictive Modeling Approach

Model Performance

PREDICTIVE MODELING APPROACH

```
# Train models for semifinal winners (excluding national champion predictions)
X_semi = df.drop(columns=['BracketEntryId', 'SemifinalWinner_East_West', 'SemifinalWinner_South_Midwest', 'NationalChampion'])
y_semi_east_west = df['SemifinalWinner_East_West']
y_semi_south_midwest = df['SemifinalWinner_South_Midwest']

model_semi_east_west = xgb.XGBClassifier(n_estimators=1500, learning_rate=0.005, max_depth=6, subsample=0.8, colsample_bytree=0.8)
model_semi_east_west.fit(X_semi, y_semi_east_west)

model_semi_south_midwest = xgb.XGBClassifier(n_estimators=1500, learning_rate=0.005, max_depth=6, subsample=0.8, colsample_bytree=0.8)
model_semi_south_midwest.fit(X_semi, y_semi_south_midwest)

# Predict the national champion, including semifinal results as additional features
X_national = df.drop(columns=['BracketEntryId', 'NationalChampion']).copy()
X_national['SemifinalWinner_East_West'] = y_semi_east_west
X_national['SemifinalWinner_South_Midwest'] = y_semi_south_midwest
y_national = df['NationalChampion']
```

MODEL PERFORMANCE

{'Semifinal East-West': {'Accuracy': 0.6895119418483905},
'Semifinal South-Midwest': {'Accuracy': 0.6432444905965156},
'Final National Champion': {'Accuracy': 0.4696357832391062}**}**

6	▲ 5	Win and Retire			0.63051

CCAC

PART 4

Key Findings & Insights

Do prediction trends change over time?

Is there significant fan bias in predictions?

Do fan predictions reflect actual team performance?

Does home vs. away performance affect predictions?

GEOSPATIAL CLUSTERING & REGIONAL VOTING TRENDS IN FAN PREDICTIONS

K-MEANS CLUSTERING ANALYSIS OF CUSTOMER GEOGRAPHIC DISTRIBUTION

CustomerDMACode	NationalChampion	MostPopularTeam	UserCount
501.0		164	5123
602.0		164	4952
652.0		164	4128
527.0		164	3822
613.0		164	3776
...		...	...
626.0		288	15
802.0		164	13
747.0		164	9
552.0		164	8
745.0		164	7

[209 rows x 2 columns]

📌 **1. Processing CustomerDMACode (DMA Code)**

Converted to a categorical variable

Calculated historical success rate for each DMA

Split CustomerDMADescription column and One-Hot Encoding

📌 **2. Processing Geographic Information (Latitude & Longitude)**

Converted to numeric format

Used mean imputation handling missing values

📌 **3. K-Means Clustering Analysis of Geographic Distribution**

Applied K-Means clustering to group customers by location

Visualized clustering results

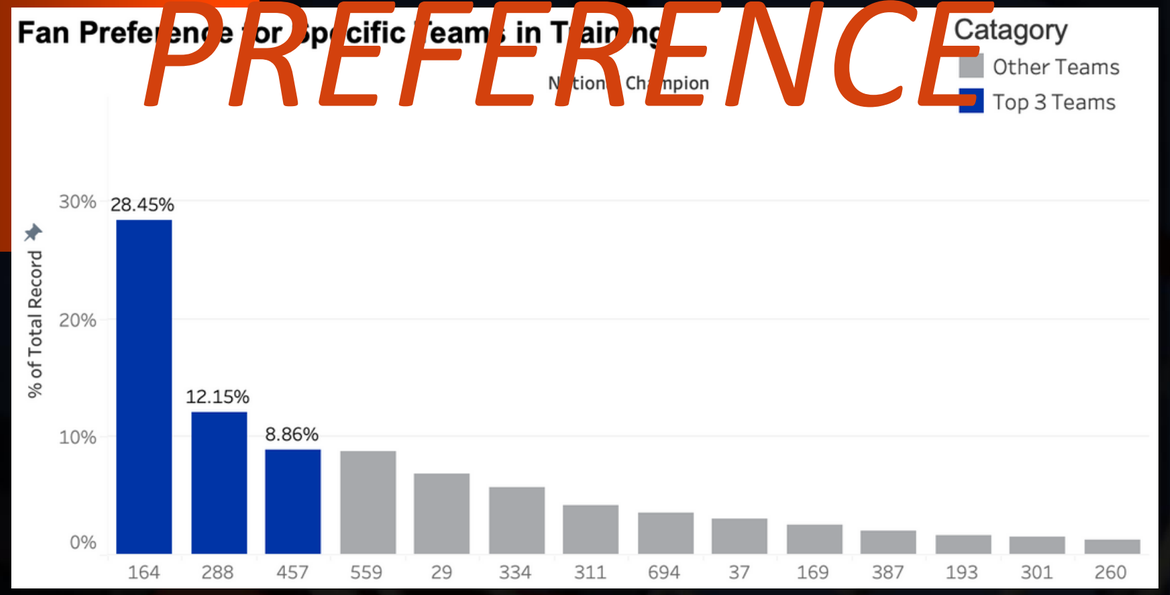
📌 **4. Analyzing Voting Trends by DMA Code**

Computed voting distribution by DMA code

Identified the most popular team in each DMA

Counted the number of users per DMA

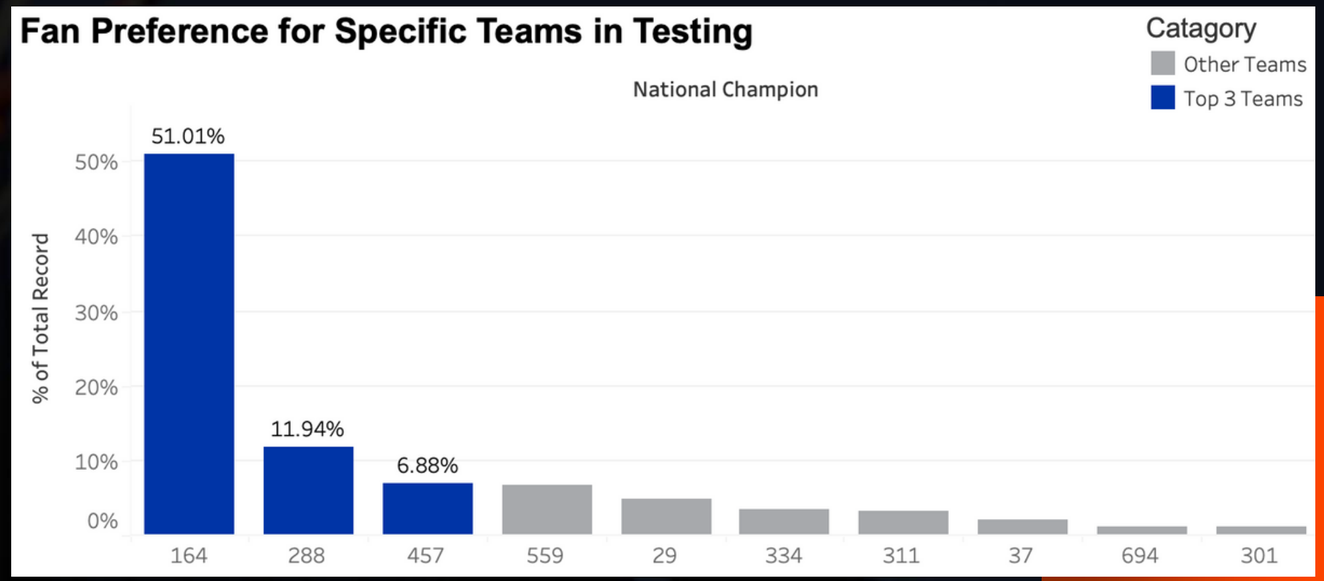
FAN PREFERENCE



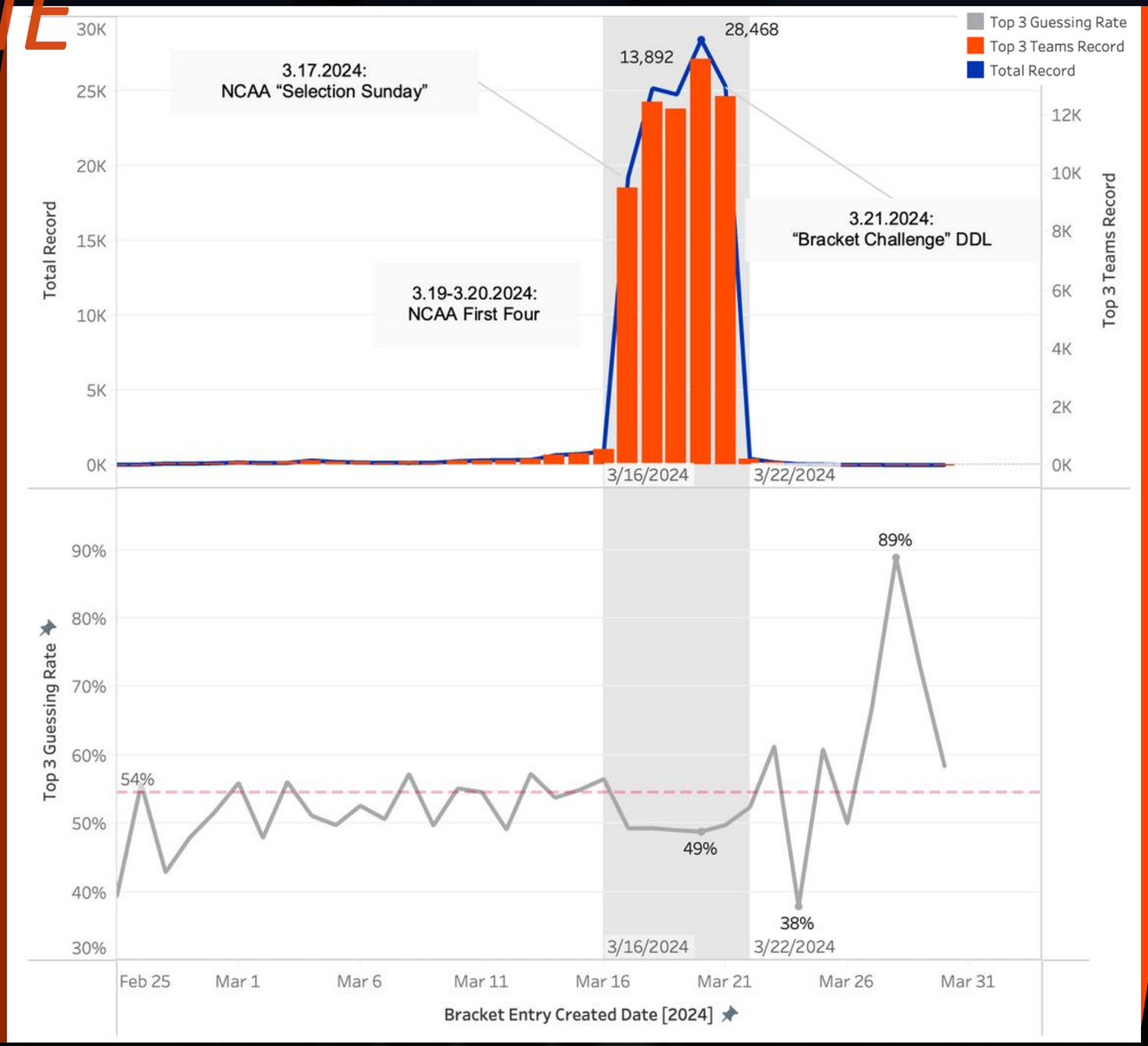
" For example, **Team 164** holds the highest percentage in both datasets, and the top five teams remain consistent across both sets.
This indicates that the model effectively captures fan preference patterns. "

" The chart presents fan prediction data (**Training**) and model-simulated fan prediction data (**Testing**) for the D-I Man Basketball National Champion.

The model successfully replicates the overall trend of fan preferences in terms of proportions. "



THE TREND IN FAN PREDICTIONS OVER TIME

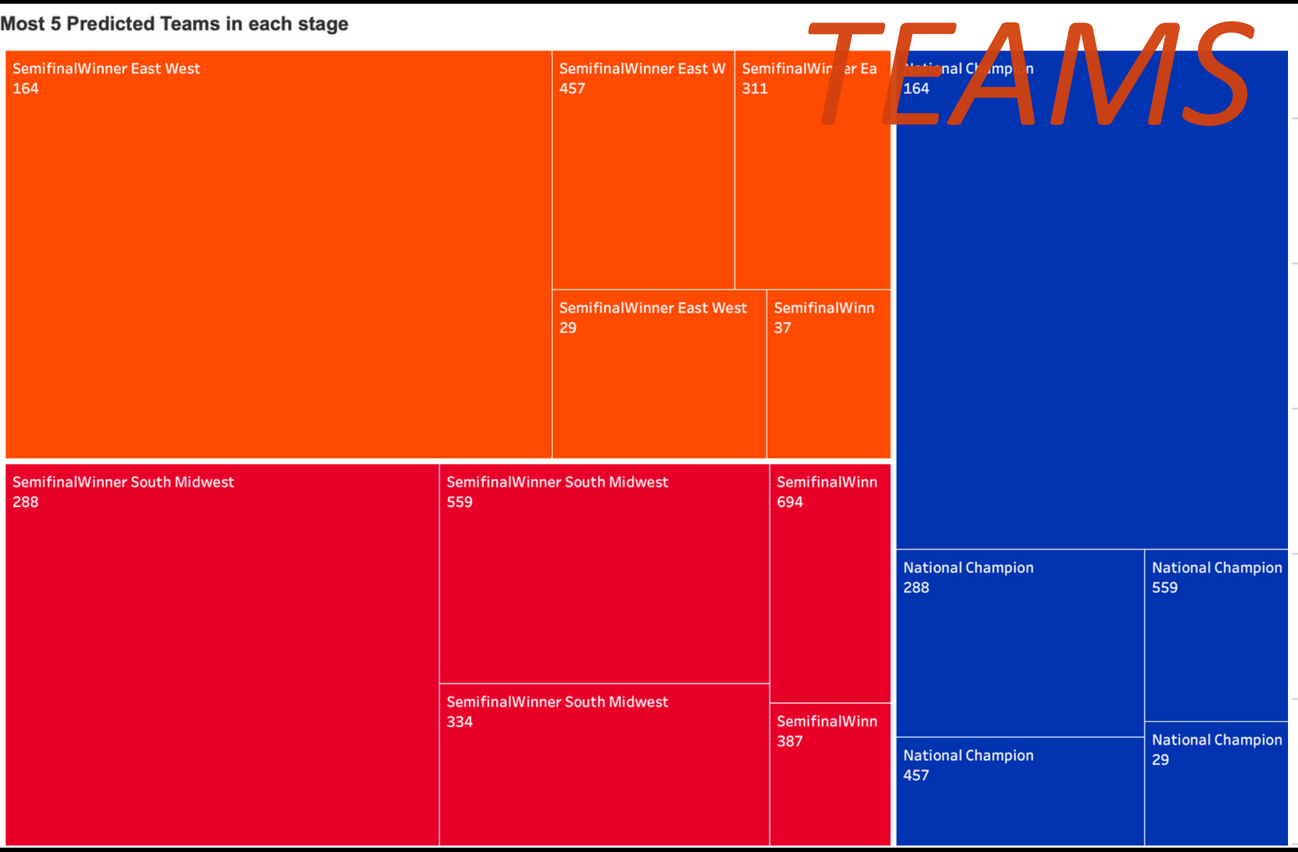


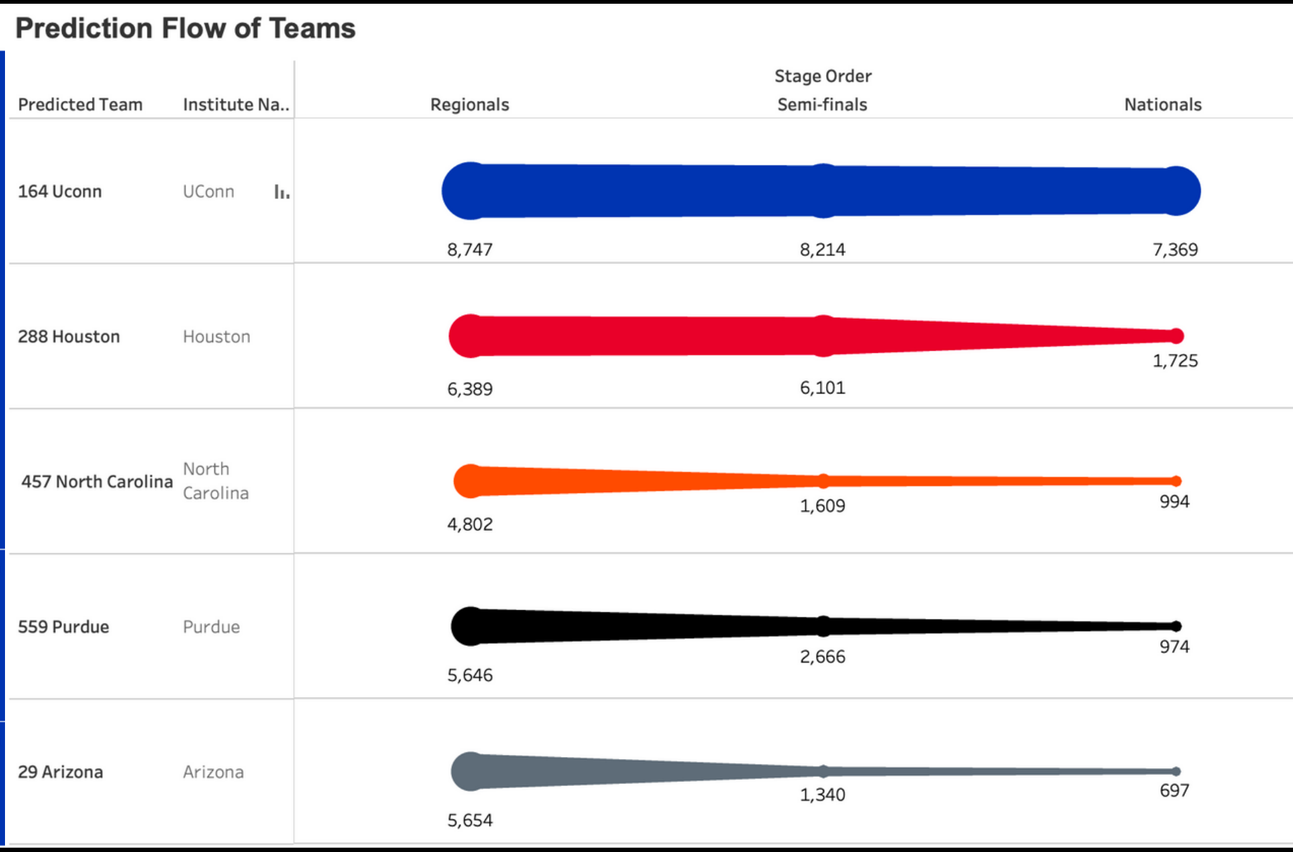
📈 **March 16–21, 2024:** Fan predictions surged to a peak, driven by three key events.

🔄 **Peak Period Trend:** The proportion of predictions for the top three teams declined, indicating fans explored more diversified strategies before the deadline.

🎯 **Post-Deadline (After March 22):** Predictions became more concentrated on the top three teams, likely influenced by actual game results or media coverage.

MOST 5 PREDICTED TEAMS





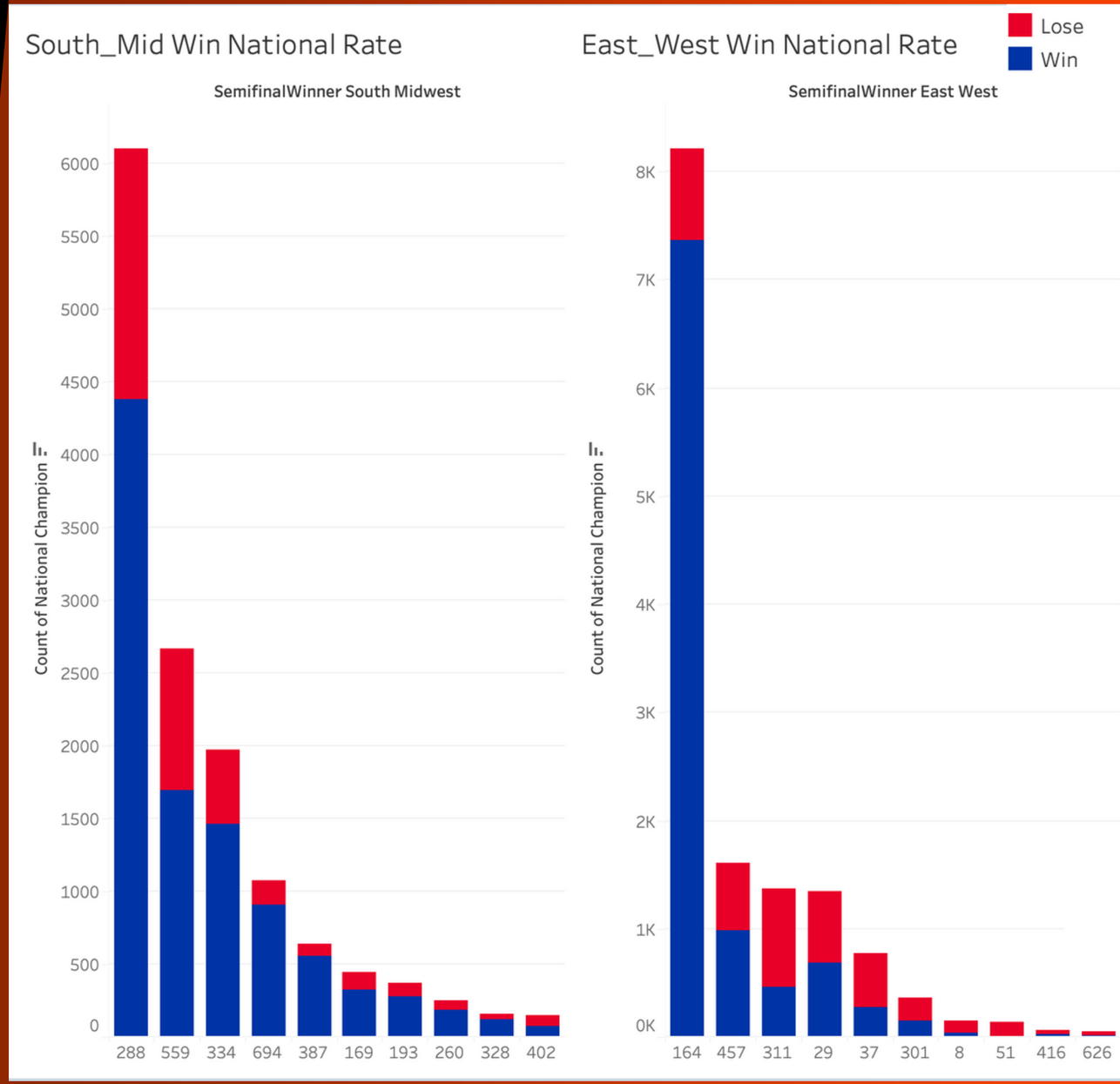
✓ **UConn** is the **most popular prediction** among fans, with a relatively stable support rate across all stages.

✓ **Houston** and **North Carolina** experienced a **significant decline** in prediction support as they progressed, possibly influenced by actual game results, team performance, or external factors such as injuries and schedule difficulty.

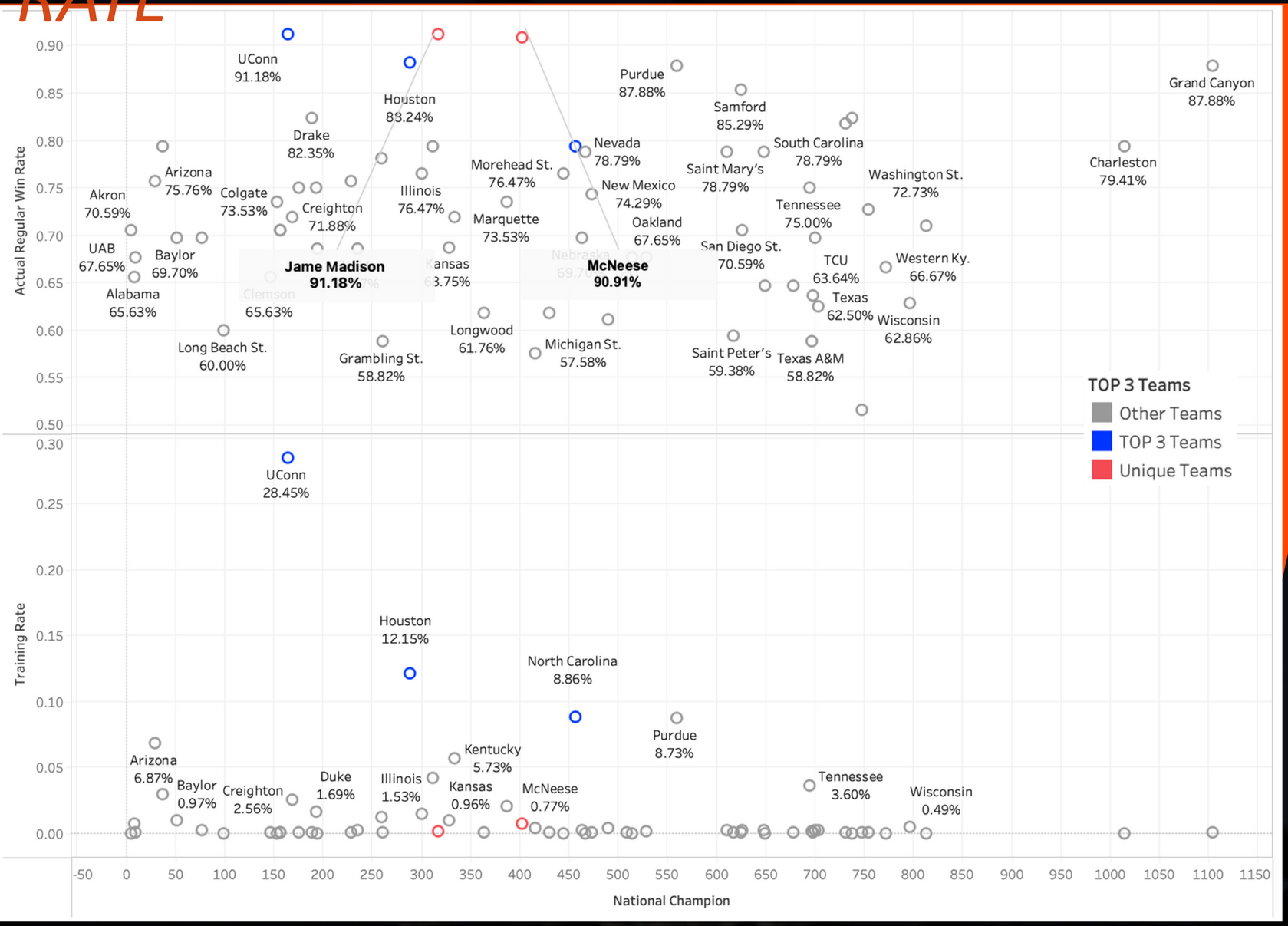
COMPARING NATIONAL CHAMPIONSHIP WIN RATES:

✓ **UConn (164)** and **Houston (288)** are the top favorites in their respective regions, having the highest number of national championships.

✓ **Teams 559** and **334** from the South Midwest region, as well as **Teams 457** and 311 from the East West region, also demonstrate strong championship competitiveness.



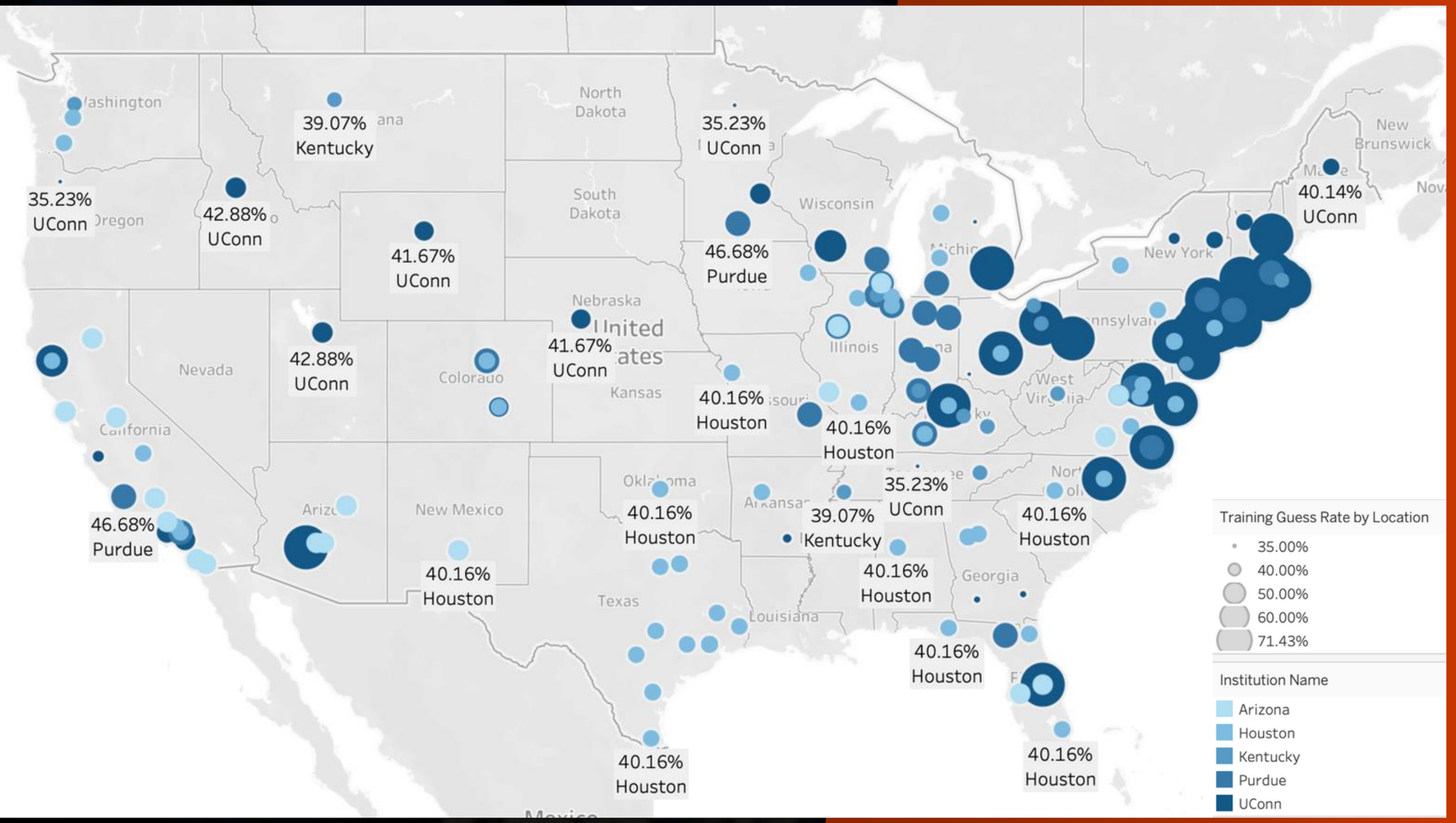
GUESSING NATIONAL CHAMPION RATE VS. REGULAR WIN RATE



✓ **Fan prediction preferences** do not always align with **actual win rates**, indicating that predictions may be influenced by **Team effect**, **historical performance**, and **emotional bias**.

✓ Some **high-win-rate teams**, such as **James Madison** and **McNeese**, are significantly **underrated** and may deserve **more attention**.

HOMETOWN BIAS OR TRUE CONTENDERS?

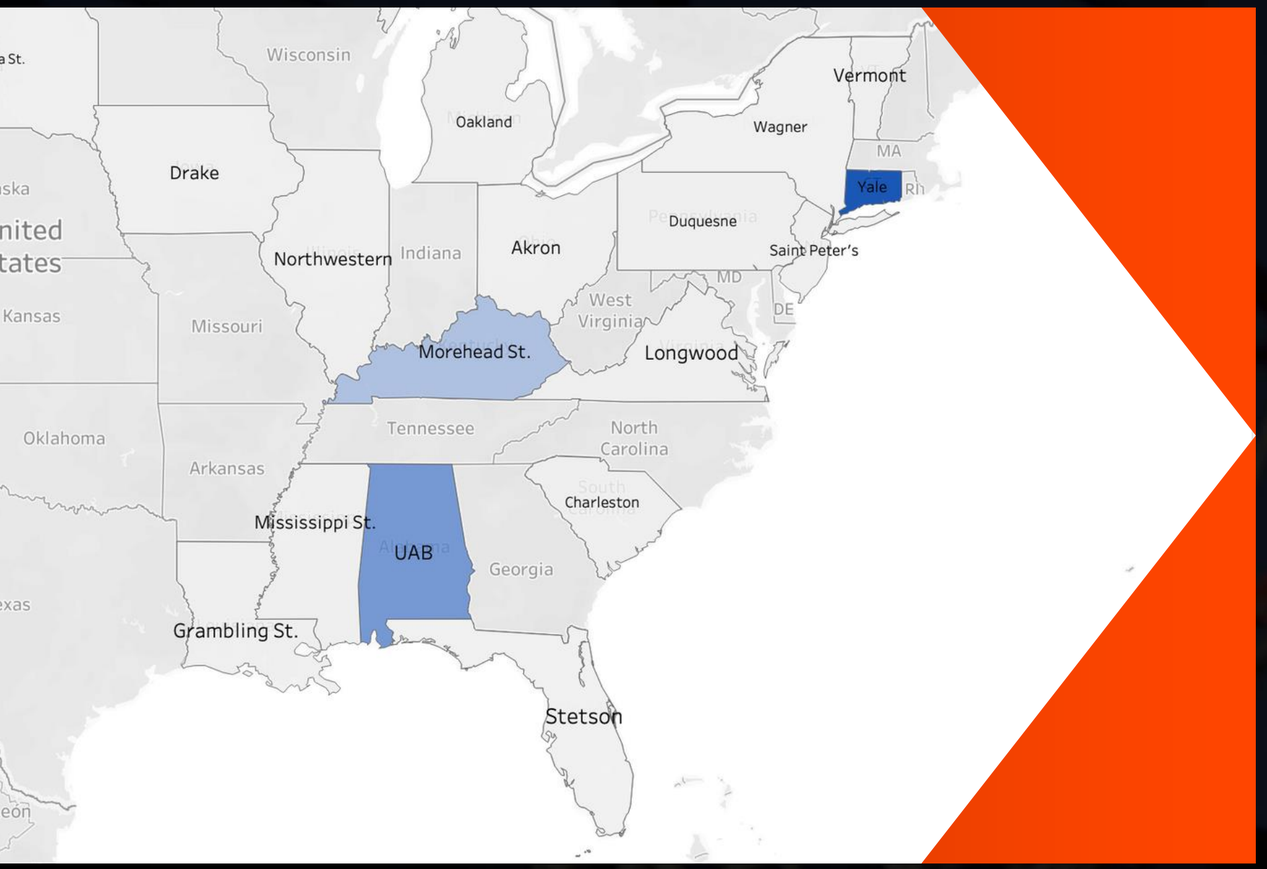


📍 **Geospatial Visualization:**

- Shows **fan support distribution** across U.S. regions.
- Bubble size = **prediction rate**.

🔍 **Key Insights:**

- **Regional Bias:** Fans favor local teams (e.g., **Houston 40.16%** in Texas, **UConn 42.88%** in Northeast).
- **Varied Preferences:** Some states show strong local loyalty, others more diverse guesses.
- **Overestimation:** Teams like **Purdue (46.66%)** in the **Midwest** may be **overrated due to hometown bias.**



REGIONAL BIAS

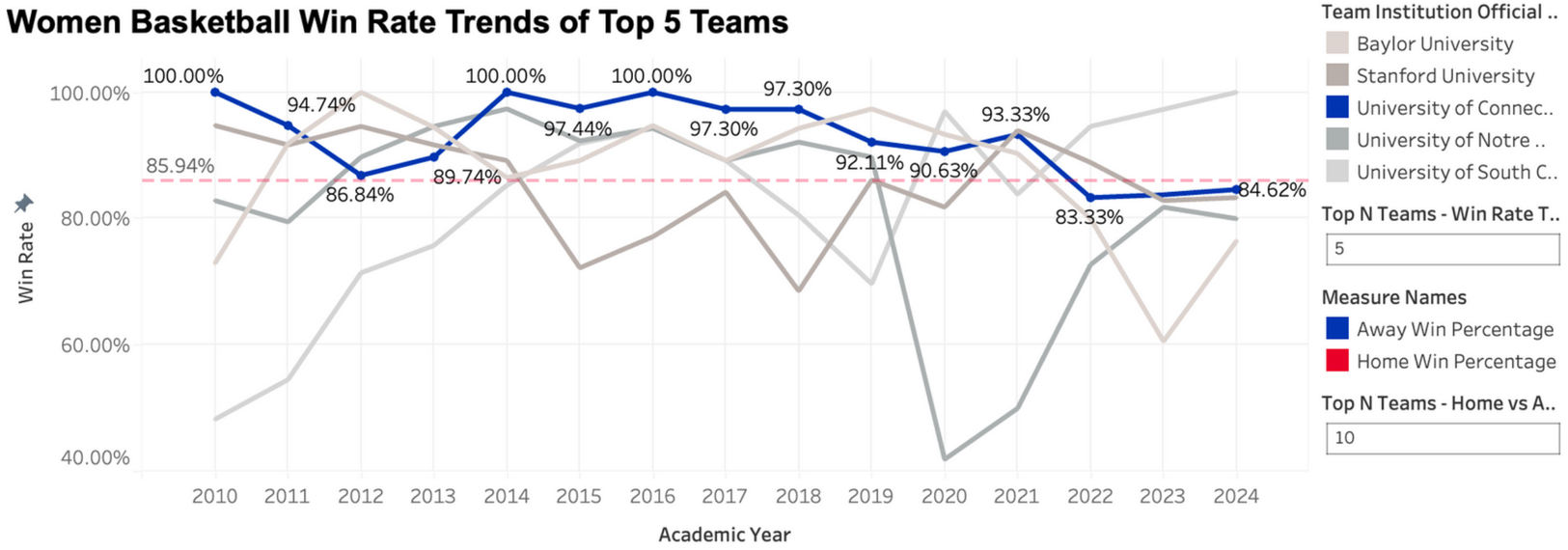
Favouritism of Predicted Team in Home State vs All States

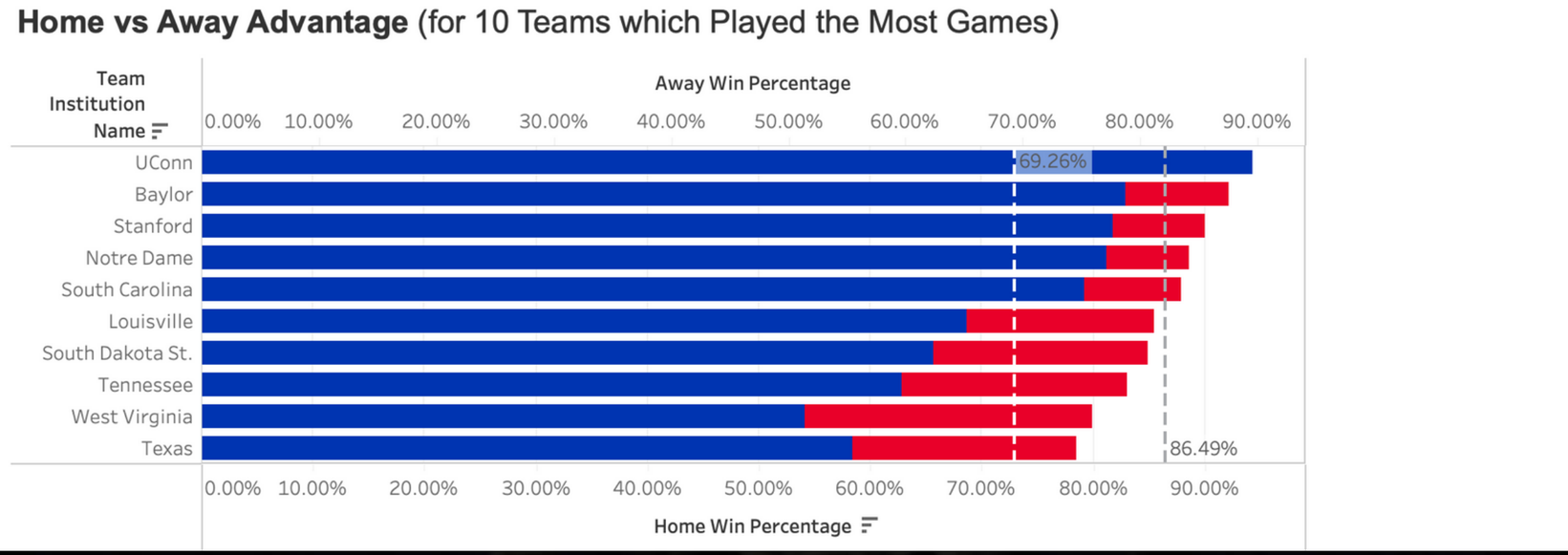
✓ **Higher Bias in Some States: States like** *Connecticut (Yale), Kentucky (Morehead St.), and Mississippi (Mississippi St.)* **show strong home-state favoritism, as seen by their darker shades.**

✓ **Small Schools Can Have High Bias: Despite not being traditional powerhouses, teams like Morehead St. and Saint Peter's have high local support, suggesting school loyalty influences predictions more than objective rankings.**

✓ **Potential School Affinity Effect: The darker the shade, the stronger the local prediction bias, suggesting that users tend to pick teams from their own state regardless of team strength.**

DOES WOMEN'S BASKETBALL POPULARITY INFLUENCE MEN'S BASKETBALL SUPPORT?





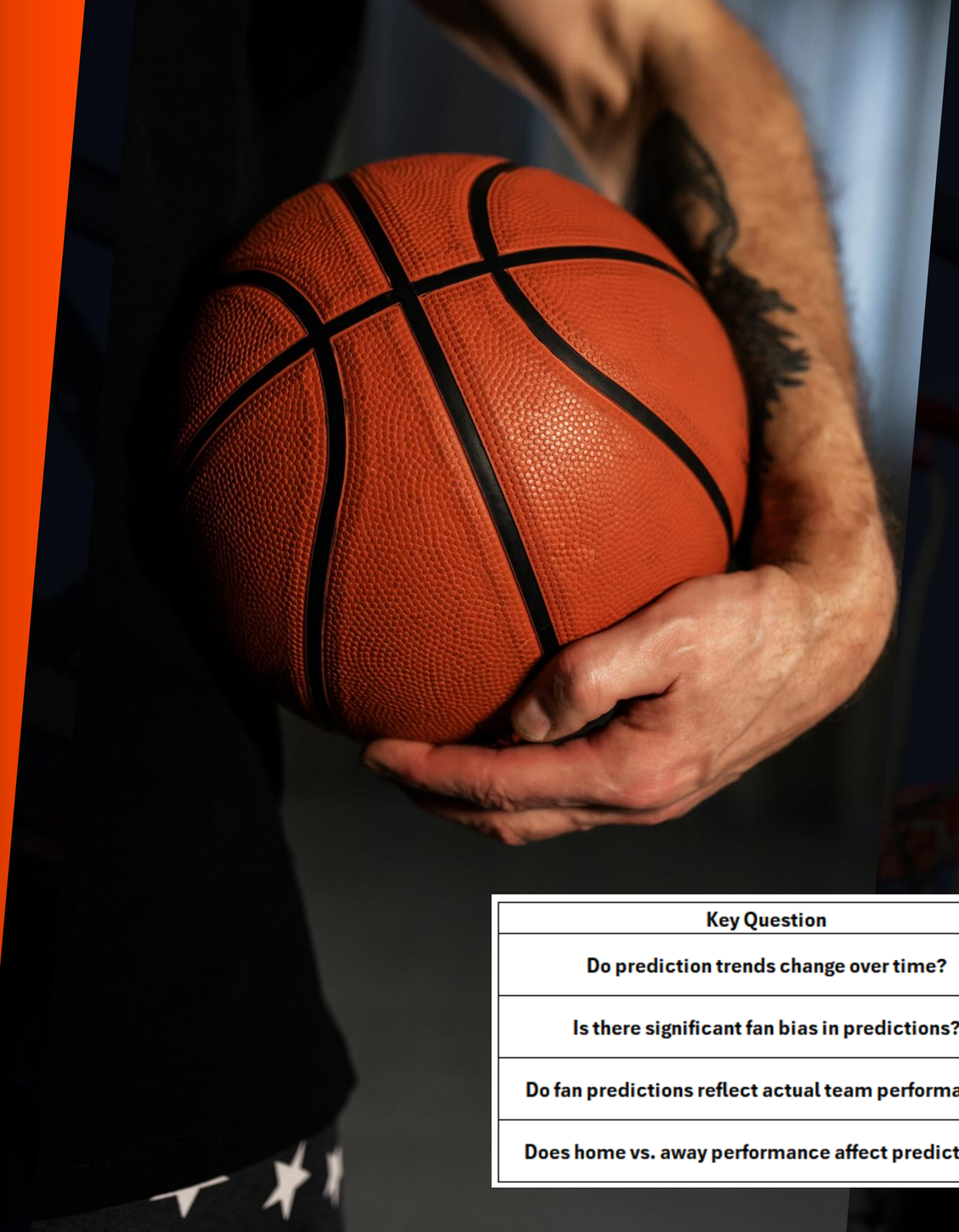
- **Strong Women's Basketball Programs → Loyal Fanbase**
  - **UConn, Stanford, and South Carolina show consistently high win rates & strong fan engagement.**
- **Possible Link Between Women's & Men's Support**
  - **Universities with dominant women's teams often have well-recognized men's programs, suggesting a potential crossover effect.**
- **Regional & Historical Influence**
  - **Fan support may extend across genders due to school-wide sports culture and local loyalty.**

PART 5

Conclusion & Future Work

Key Takeaways

Future Work



KEY TAKEAWAYS

1 Predictive Model Performance

📌 Kaggle Leaderboard Performance

Our model achieved a Kaggle score of 0.63051.

2 School Affinity & Fan Bias Findings

📌 Key Observations on Fan Loyalty & Bias

- Fans exhibit strong school affinity, favoring local teams over objectively stronger competitors.
- Regional clustering reveals distinct voting biases—certain states heavily support their home institutions, even when their teams are underdogs.
- Historical school performance influences both men's and women's basketball predictions, reinforcing the idea that fans are loyal to the school, not just a specific team.

3 Additional Analytical Findings (Part 4 Insights & Extended Analysis)

📌 Exploring Trends in Fan Predictions

Key Question	Key Insight
Do prediction trends change over time?	Prediction trends fluctuate, especially before major tournament deadlines. Fan choices diversify as deadlines approach.
Is there significant fan bias in predictions?	Regional and school affinity biases are evident; fans favor local teams and historically strong programs.
Do fan predictions reflect actual team performance?	Clear mismatch between fan selections and actual team strength; historical reputation often outweighs real-time performance.
Does home vs. away performance affect predictions?	Fans favor teams with strong home-court records, despite NCAA games being played on neutral courts.

FUTURE WORK

Address Data Bias in Predictions

- If fans over-support popular teams, the model may reinforce this bias, reducing accuracy in identifying underdog victories.
- Future improvements should include bias correction techniques such as reweighting predictions or incorporating adjusted ranking features.

Improve Feature Utilization in Predictive Models

- Although we explored One-Hot Encoding, it was not fully implemented in the predictive model.
- Future iterations should integrate richer categorical features, such as conference affiliations, historical trends, and game-level statistics, to improve prediction accuracy.

Enhance Model Interpretability & Feature Expansion

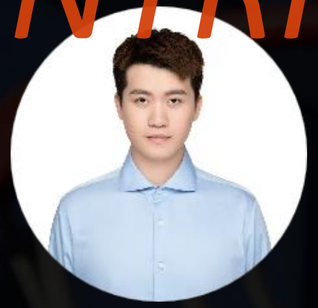
- Expand feature engineering efforts, incorporating additional external data sources to strengthen model insights.

APPENDIX

→ **Team Roles & Contributions**

→ **Acknowledgments**

TEAM ROLES & CONTRIBUTIONS



Huangkun Chen

Leader
Presentation
Tableau Analysis
Kaggle Submission

Email

chen5180@purdue.edu

Yingtong Wang

Presentation
Tableau Analysis
Kaggle Submission
Task Allocation

Email

wang6679@purdue.edu

Yingjie Chen

Tableau Analysis
Communication

Email

chen5301@purdue.edu

Nikhil Ram Atlurin

Tableau Analysis
Kaggle Submission

Email

atlurin@purdue.edu

We are all Purdue Univeristy MSBAIMer !

ACKNOWLEDGMENTS

We sincerely thank the **Crossroads Classic Analytics Challenge (CCAC) organizers** for providing us with this invaluable opportunity to apply data science techniques to real-world sports analytics. This competition has significantly enhanced our skills in predictive modeling, data analysis, and fan behavior research.

Additionally, we extend our deepest appreciation to our **team members** for their dedication, collaboration, and hard work throughout every stage of the project—from the Kaggle competition to the creation of the TABLEAU dashboard, and finally, to the video presentation. The collective effort in each phase has continuously propelled our analysis forward, and the strong synergy within our team has been key to the success of this research.

Lastly, we are grateful to the **data providers, Kaggle, and all mentors and faculty members** for offering us the NCAA bracket prediction dataset and providing invaluable knowledge support, enabling us to conduct an in-depth study on fan prediction behaviors.

CCAC

Unveiling Fan Bias in NCAA Bracket Predictions:

Trends, Accuracy, and the Impact of School Affinity

THANK YOU

Team: Win and Retired