



# ***CROSSROADS CLASSIC 2024***



A collage of three action shots from a volleyball game. The left image shows a player in a red jersey diving for a ball. The center image shows a player in a white jersey with blonde hair flying through the air. The right image shows a player in a white jersey. The background is a solid blue color with a black diagonal stripe running from the top left to the bottom right.

# ***TEAM GREEN LIGHT DISTRICT***



# MEET THE TEAM



***DURGA DASH***



***SOHAM AGARWAL***



***SOHAN SAHOO***



***VISHNU PONDURI***






***NAGARJUNA  
CHIDARALA***





**BEFORE WE SHOW  
YOU HOW WE  
RANKED #1 ON  
KAGGLE**

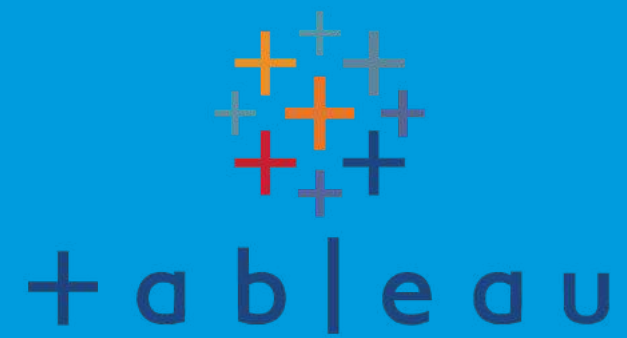
## Public Leaderboard

1	Green Light District	    	0.98795	59	6d
---	----------------------	---	---------	----	----

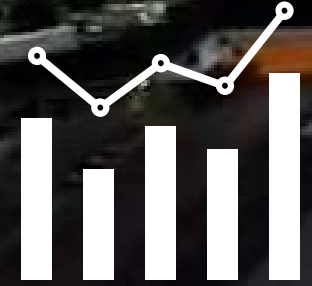
## Private Leaderboard

1	— Green Light District	    	0.98682	59	6d
---	------------------------	---	---------	----	----





***LET'S TAKE IT  
FROM THE START***



**Train Data: ~ 210k rows**  
**Test Data: ~ 21k rows**  
**Columns ~ 25**



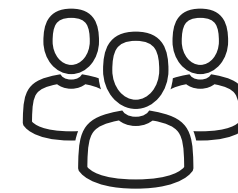
**Customer related data:**  
**Geographical data**  
**Historical data**  
**Interaction data**



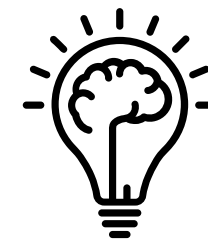
**Event-related data:**  
**Event type**  
**Geographical data**

**WHAT ARE**

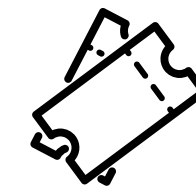
**WE TRYING  
TO SOLVE?**



**How do we increase the ticket sales for the NCAA?**



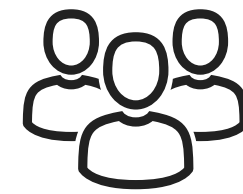
**How do we Identify potential brokers by analyzing suspicious behaviours?**



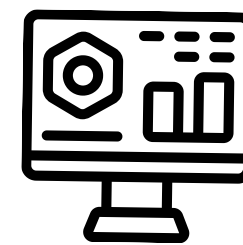
**How do we identify who buys different types of tickets?**

# WHY?

# HOW?



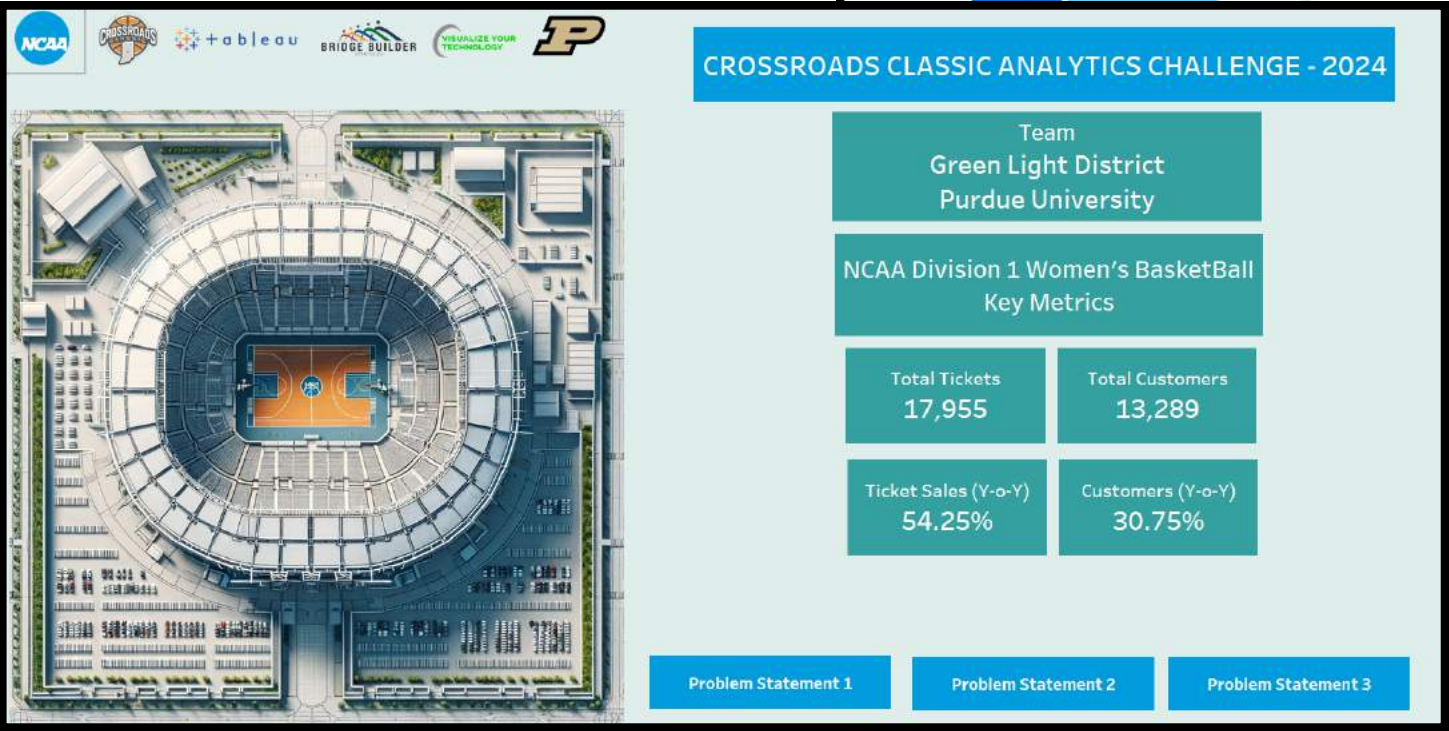
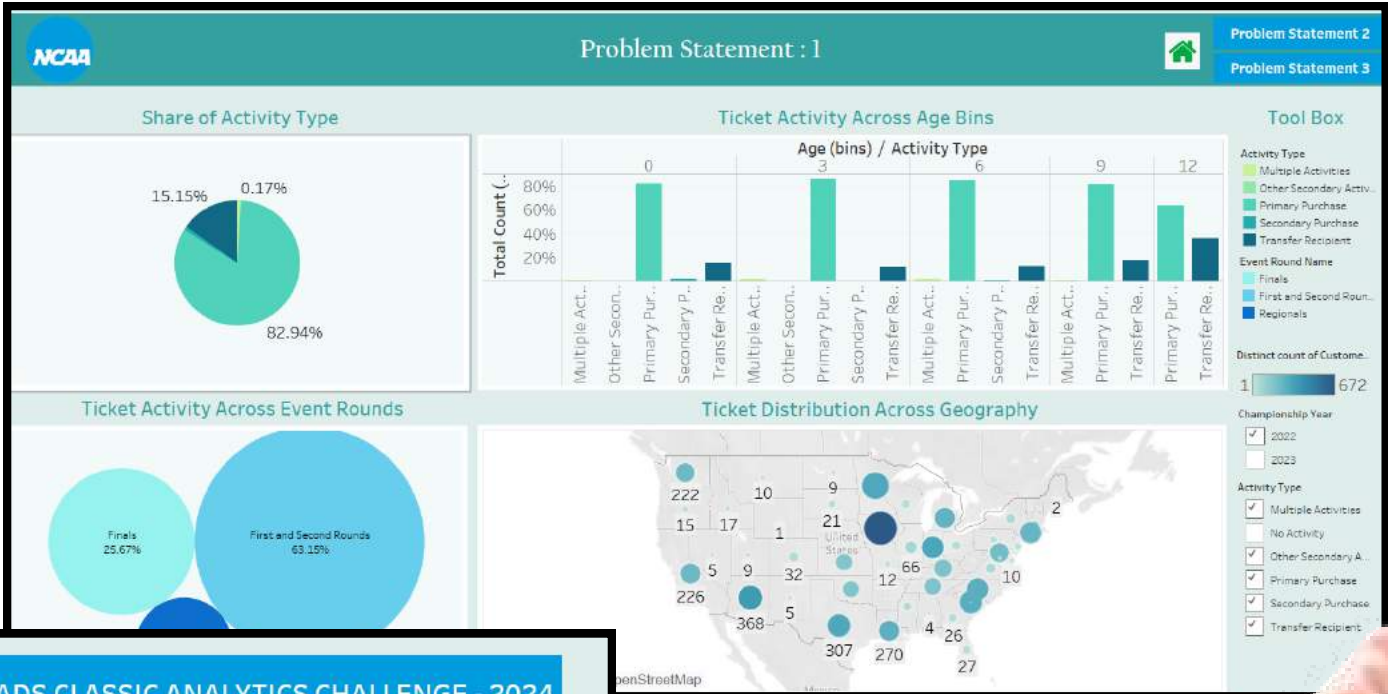
Understanding customer's behavior helps the NCAA make decisions and increase sales.



Using visualizations via Tableau and model building using Python utilizing data from the NCAA.



# OUR INTERACTIVE DASHBOARDS





# PROBLEM STATEMENT 1

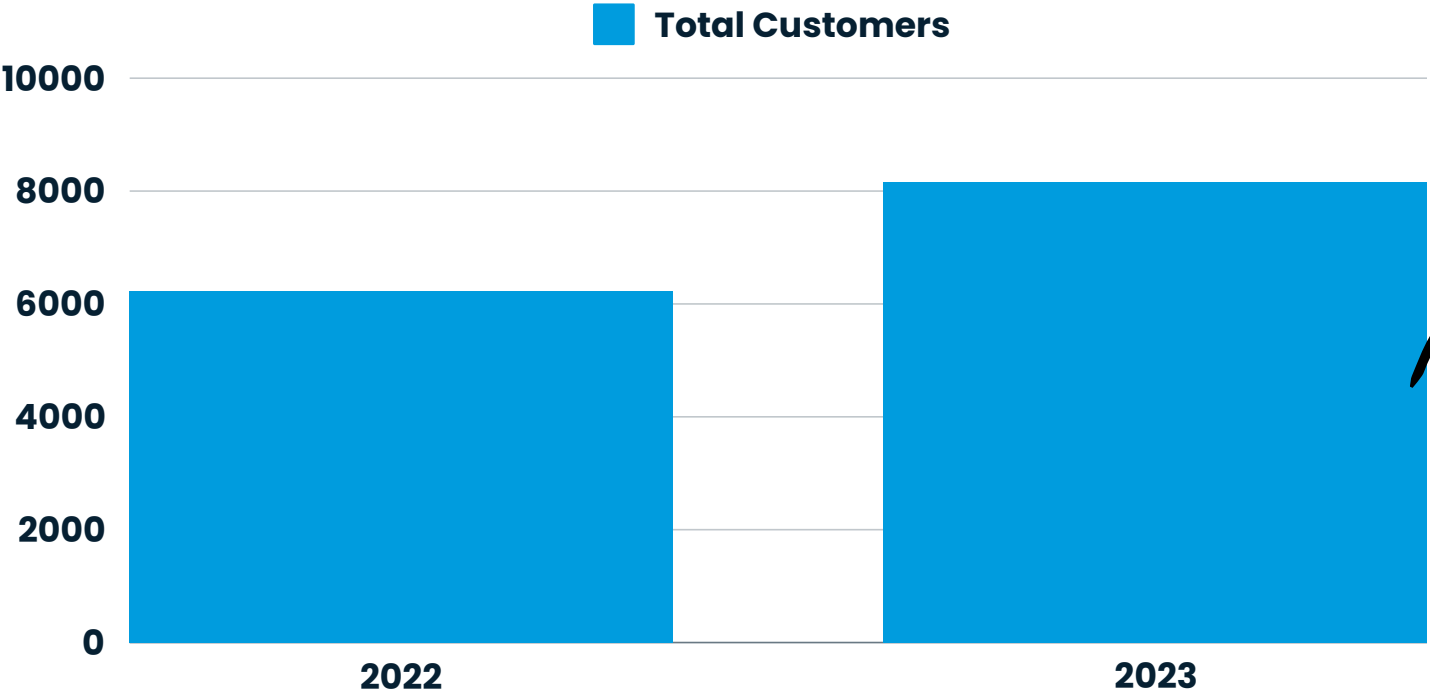
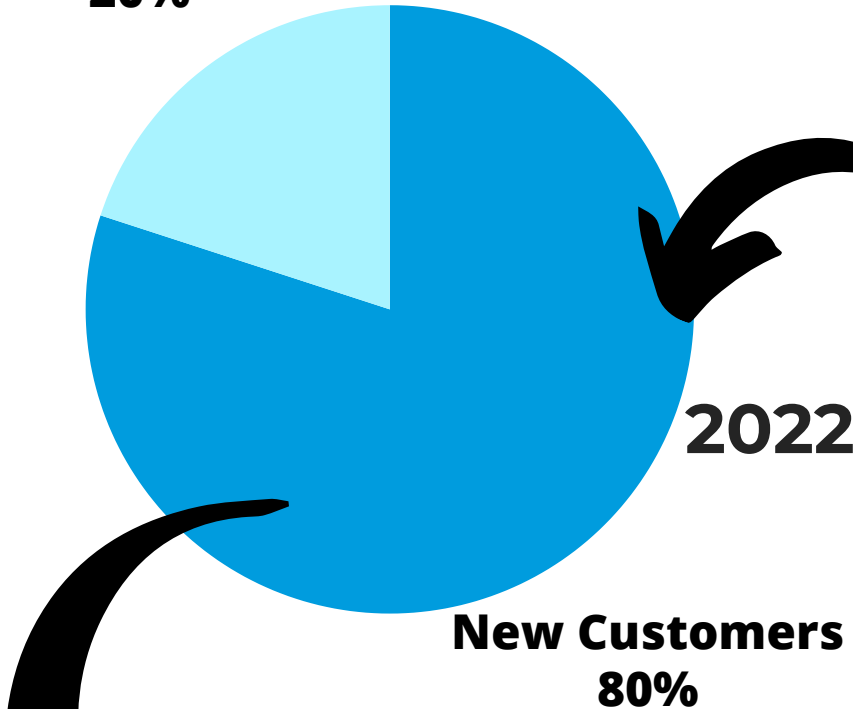


How do we increase the ticket sales?

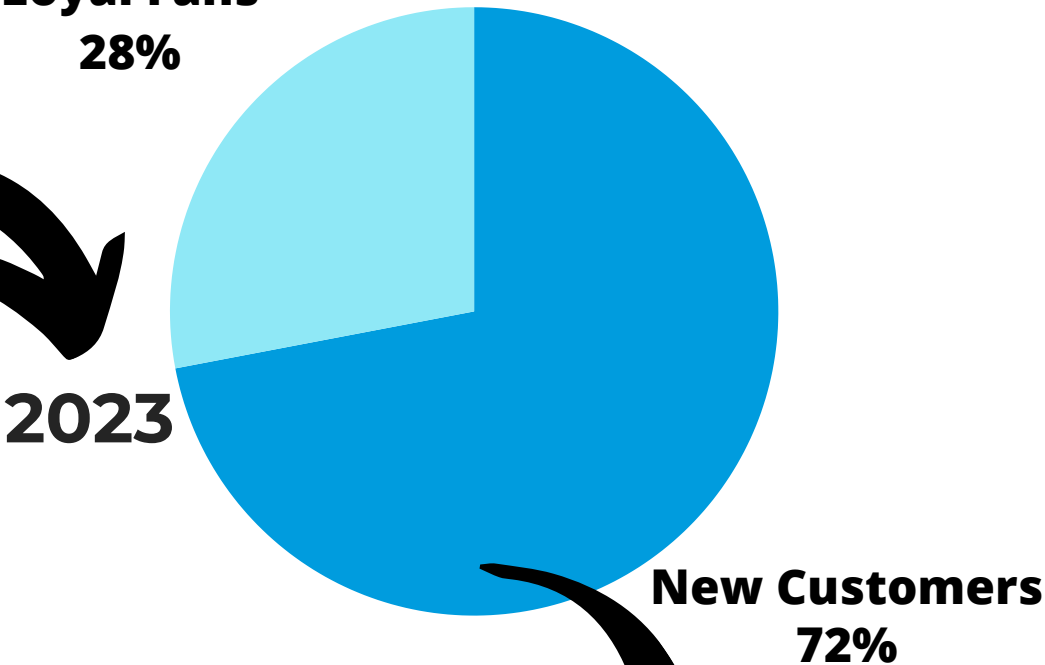
# CUSTOMER BASE SEGMENTATION



**Loyal Fans**  
**20%**



**Loyal Fans**  
**28%**

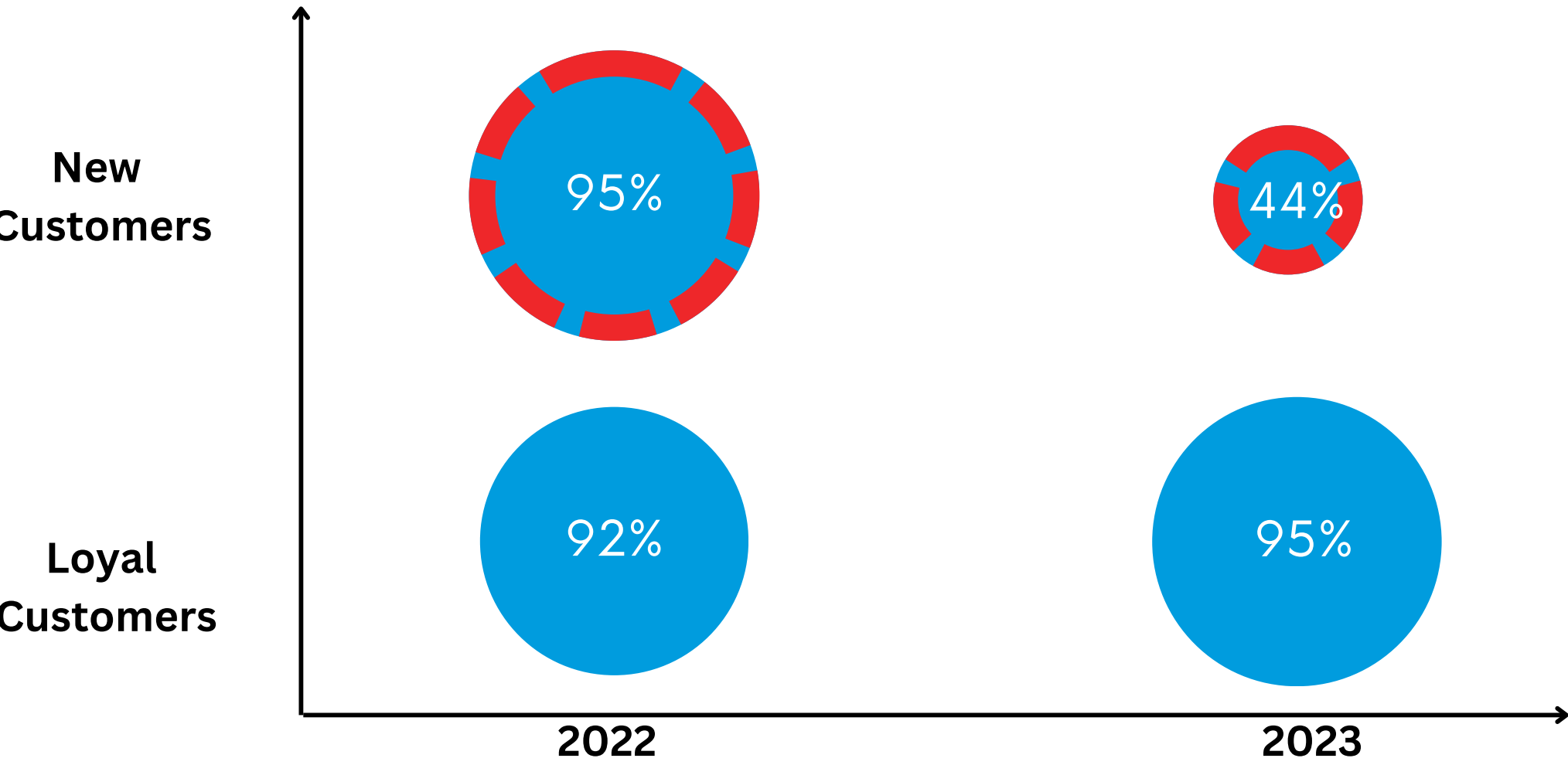


Host Type	% new customers	Median Distance
Host	72%	12.4
Non Host	28%	311

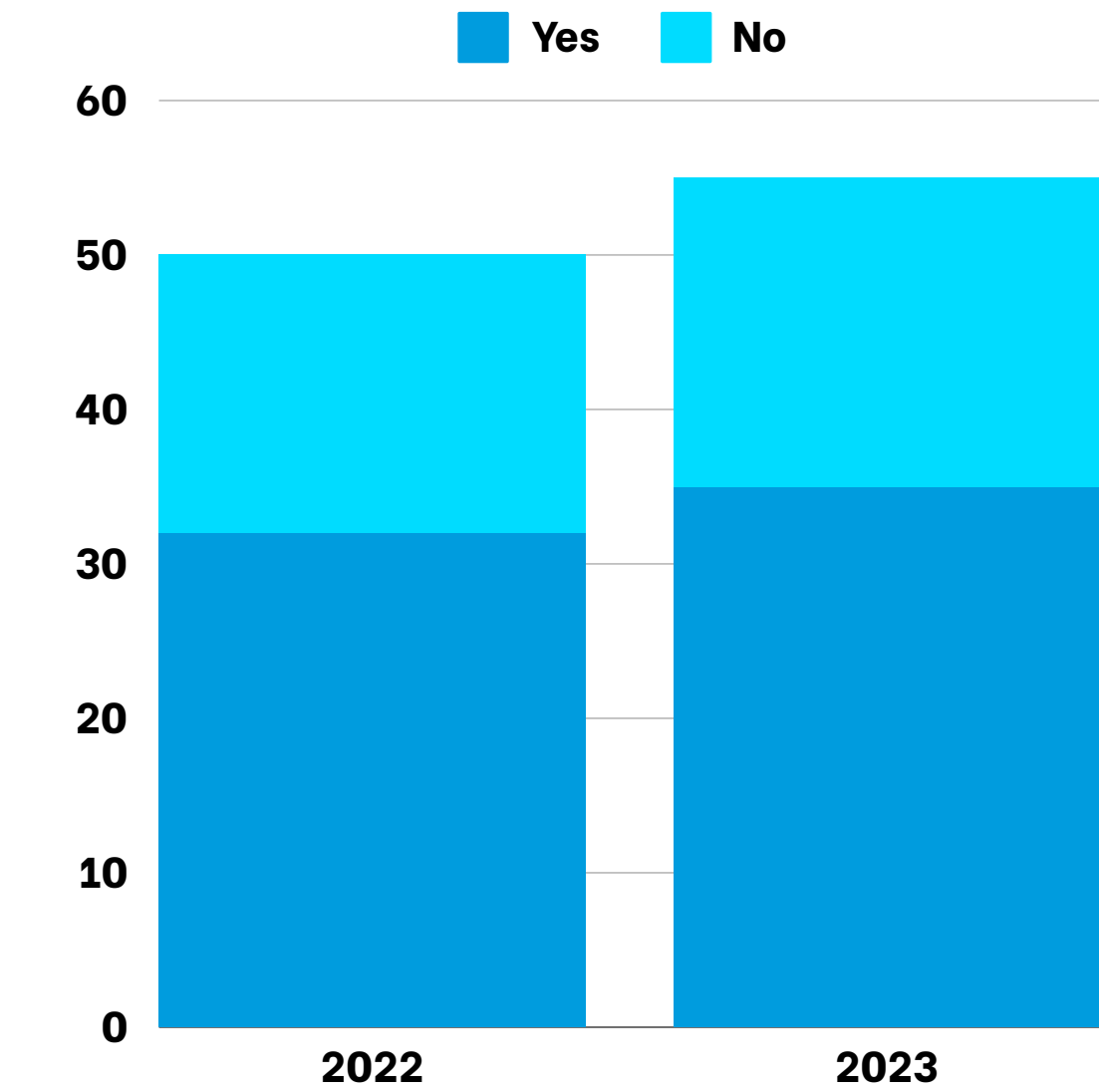
Host Type	% new customers	Median Distance
Host	68%	16
Non Host	32%	389



# ESTABLISHING A CONNECT WITH CUSTOMERS



% customers mails were sent



Mail Open Frequency by Customers

- Inconsistent trend in emails sent to new customers from 2022 to 2023
- Customers who attend are likely to open thus ensuring good communication.

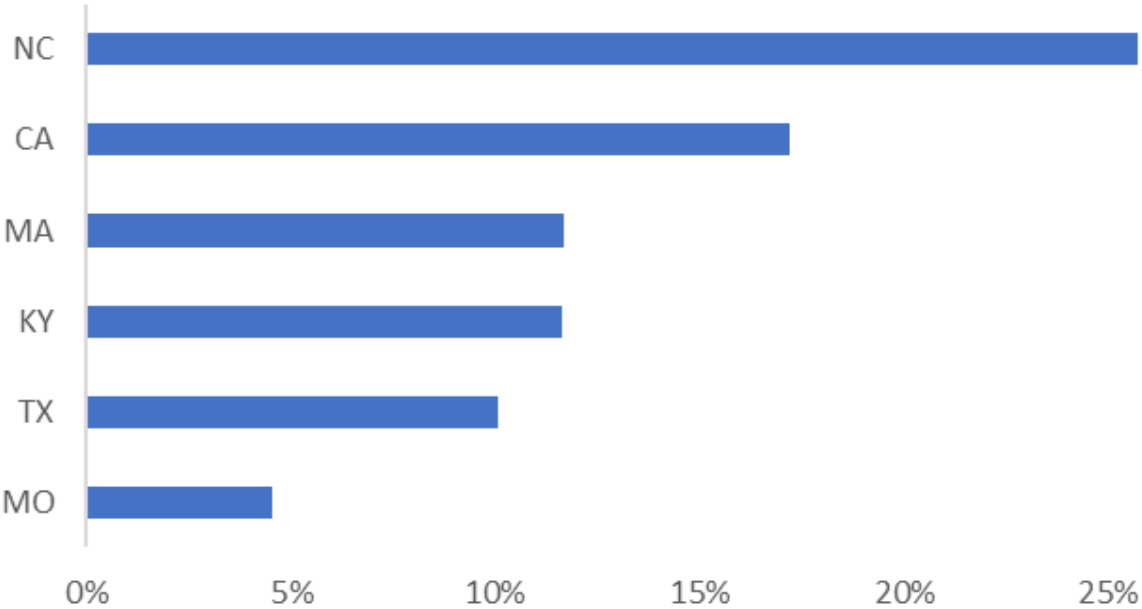
# POTENTIAL OPPORTUNITIES IN NON-HOST STATES

2022



- Certain cities despite being hosts, are not able to attract customers
- Conversion rate hover around 15% in these states
- Our model suggests location as one of the biggest features in predicting whether someone will purchase!

Bottom 5 host states with least conversion for year 2022

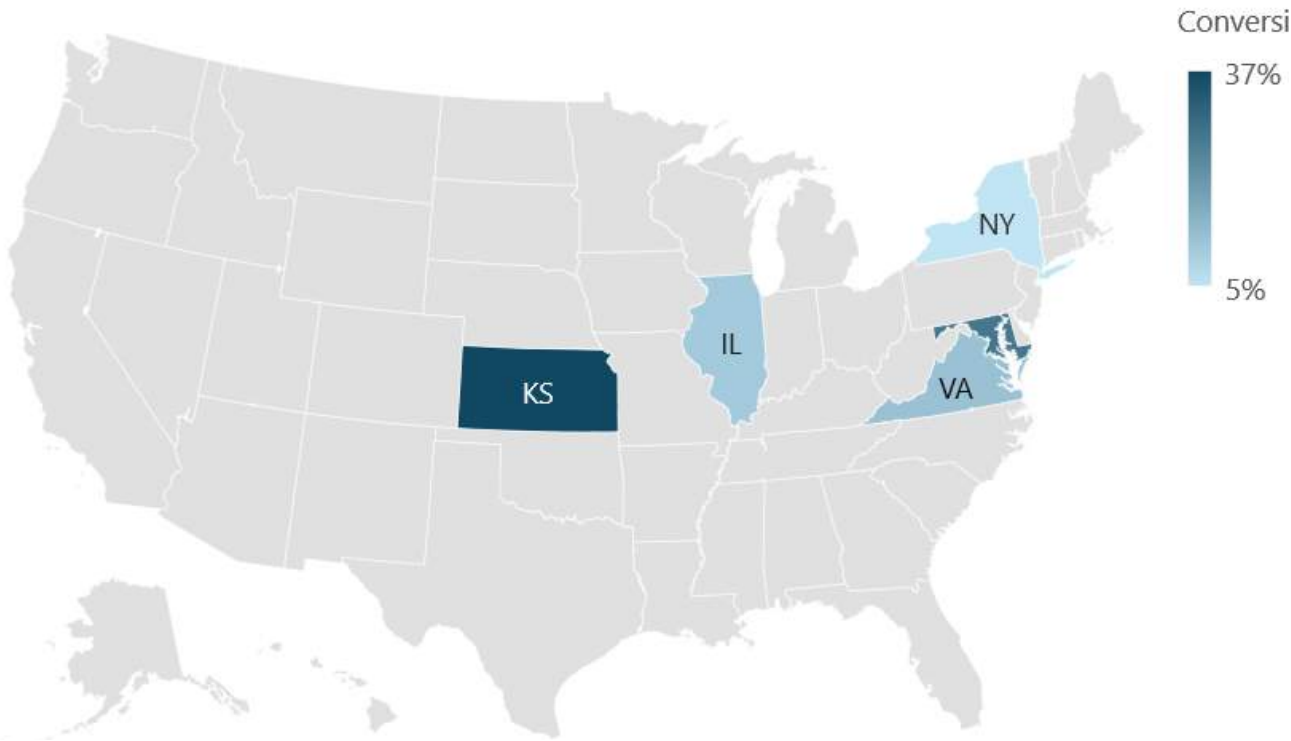


## POTENTIAL OPPORTUNITIES

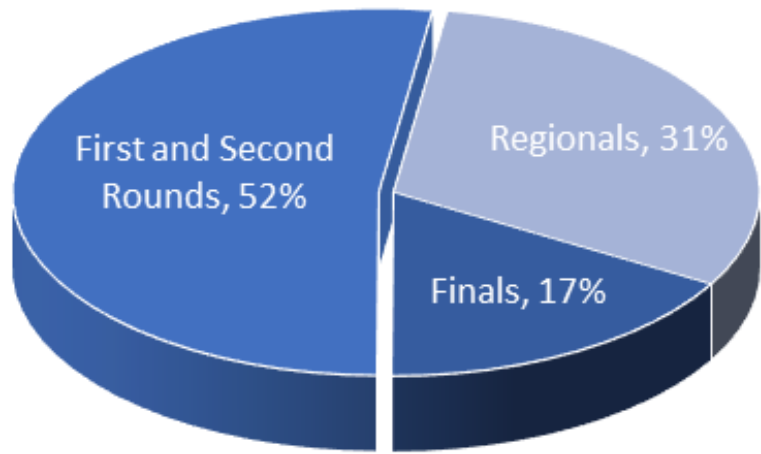
- Non host states (2022)

States	Distance	Customers	Age
MD	368	210	0.8
KS	180	162	0.21
IL	352	66	0.63
VA	388	49	0.93
NY	813	41	0.65

2022- Non Host states with max customers



Tickets distribution across event rounds from non host states in 2023



Finals First and Second Rounds Regionals

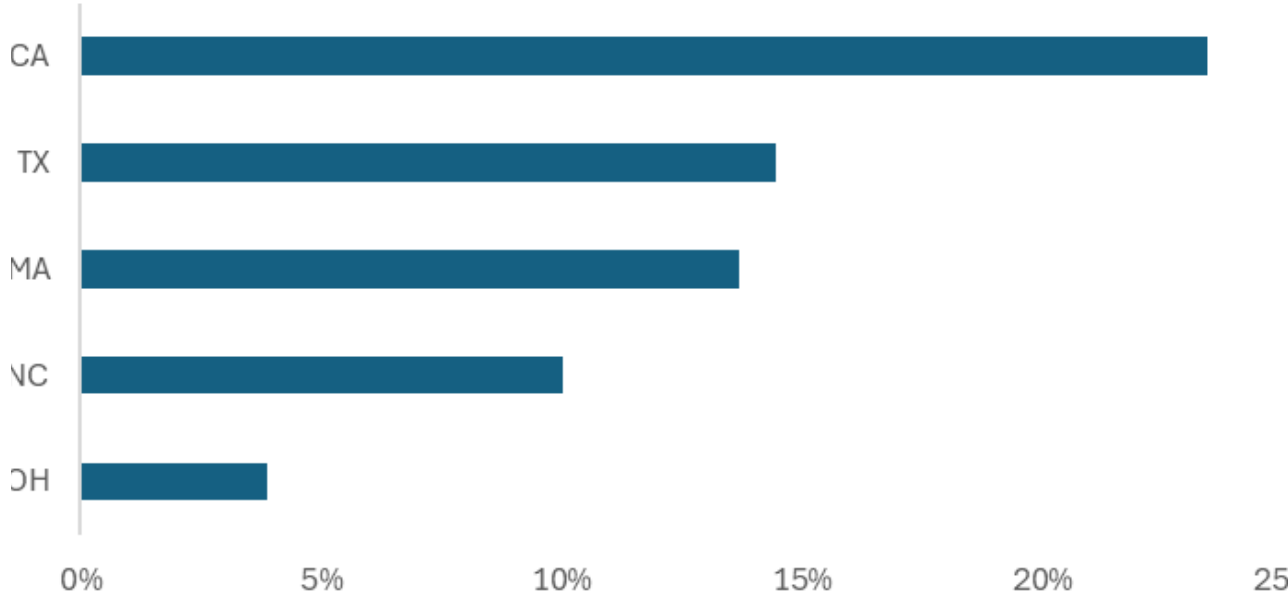
# POTENTIAL OPPORTUNITIES IN NON-HOST STATES

2023



- Certain cities despite being hosts, are not able to attract customers
- Conversion rate hover around 15% in these states
- Our model suggests location as one of the biggest features in predicting whether someone will purchase!

Bottom 5 host states with least conversion for year 2023

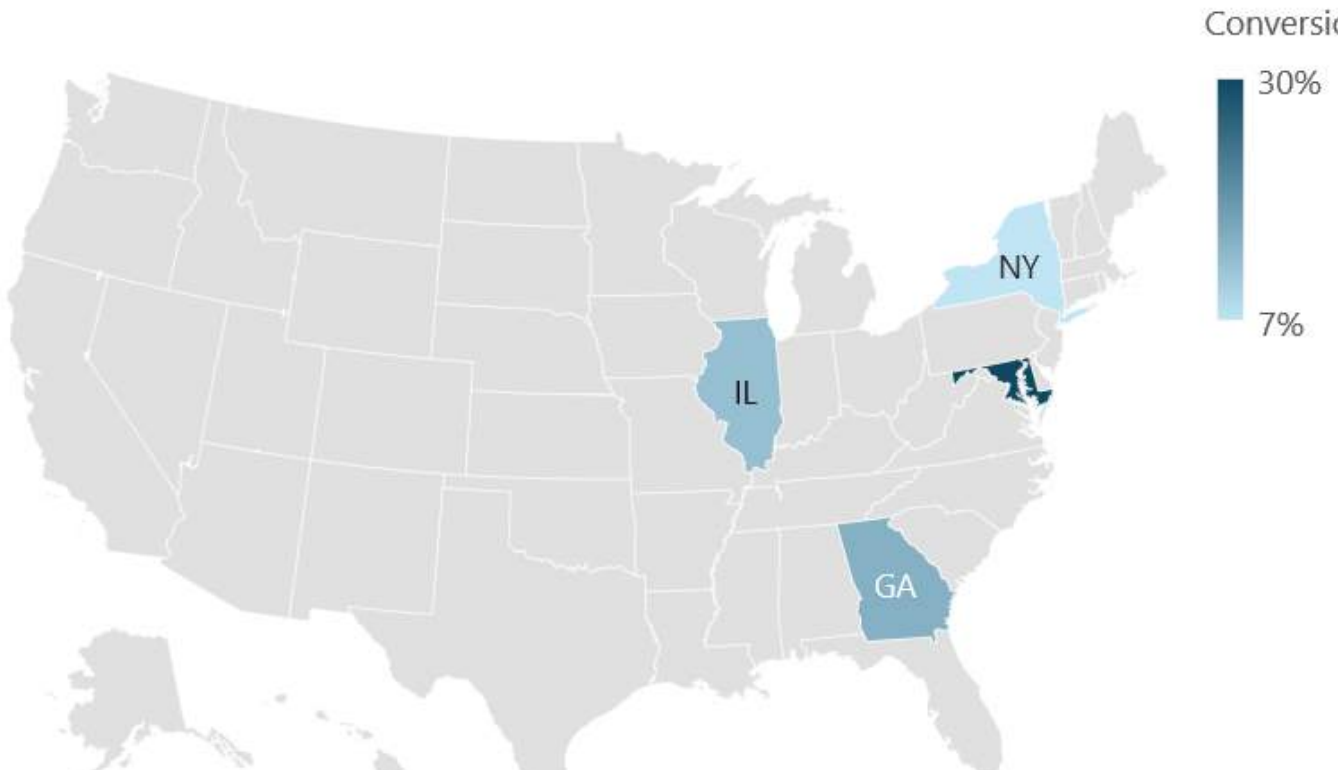


## POTENTIAL OPPORTUNITIES

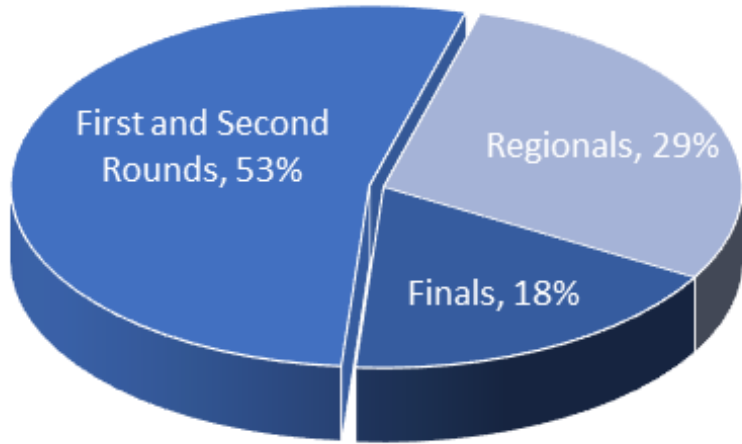
- Non host states (2023)

States	Distance	Customers	Age
MD	371	288	0.36
IL	781	91	0.25
GA	149	77	0.15
NY	1298	64	0.18
NJ	1608	53	0.2

2023- Non Host states with max customers



Tickets distribution across event rounds from non host states in 2023



Finals First and Second Rounds Regionals



# PROBLEM STATEMENT 2

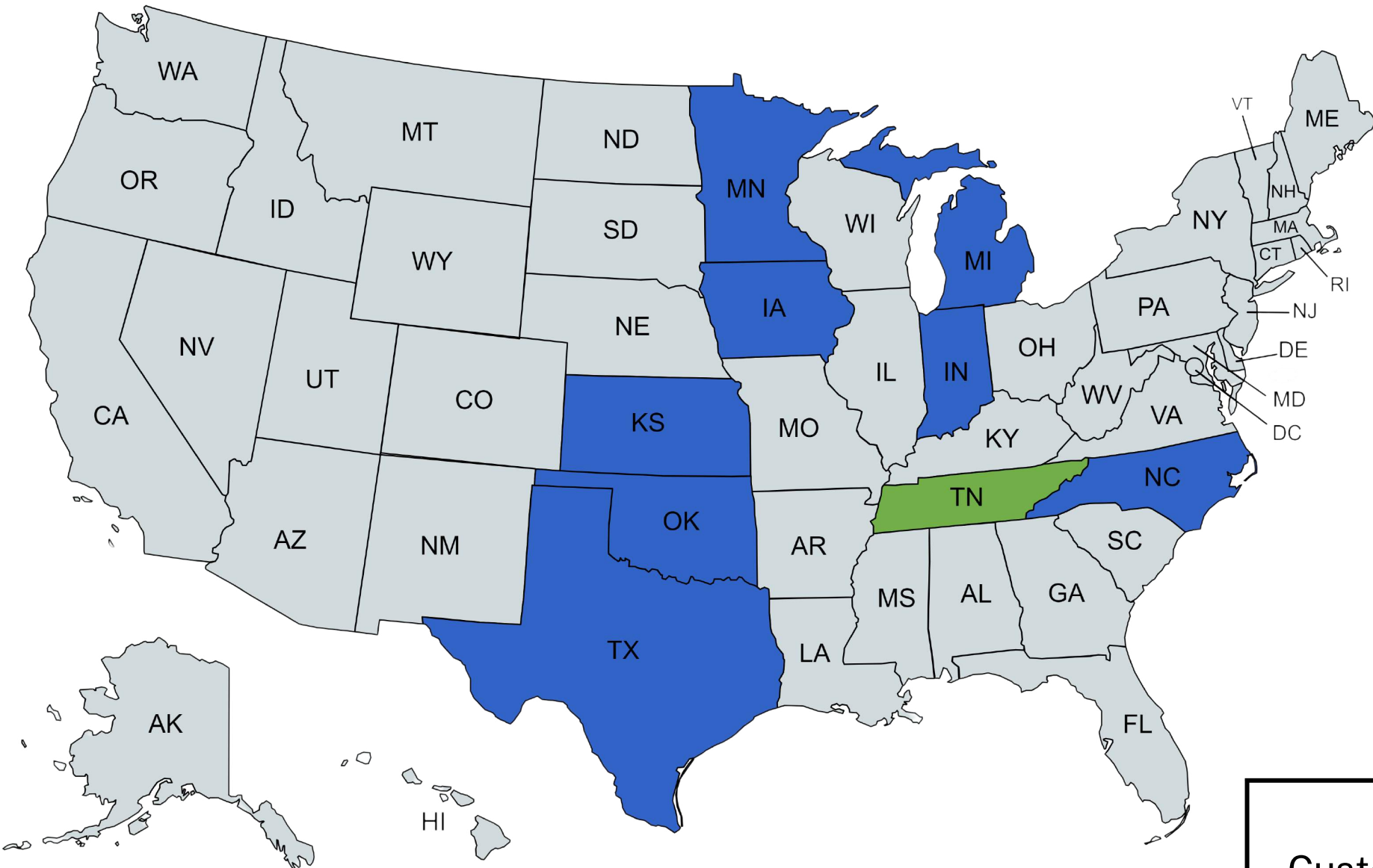


How do we identify potential brokers with unusually high purchases by analyzing suspicious behaviors?

# IDENTIFYING POTENTIAL BROKERS



2022



Championship Year: 2022  
CustomerID: 494708

Host cities  
Origin

## Methodology

- Identified distinct event dates for which a particular customers has booked tickets for
- Mapped to distinct hosting cities
- Identified ticket type preference

## For the particular customer in reference

- He/She has booked tickets to all 10 different states for 5 different events
- Made all primary purchases and has to travel 6500 miles to physically witness matches

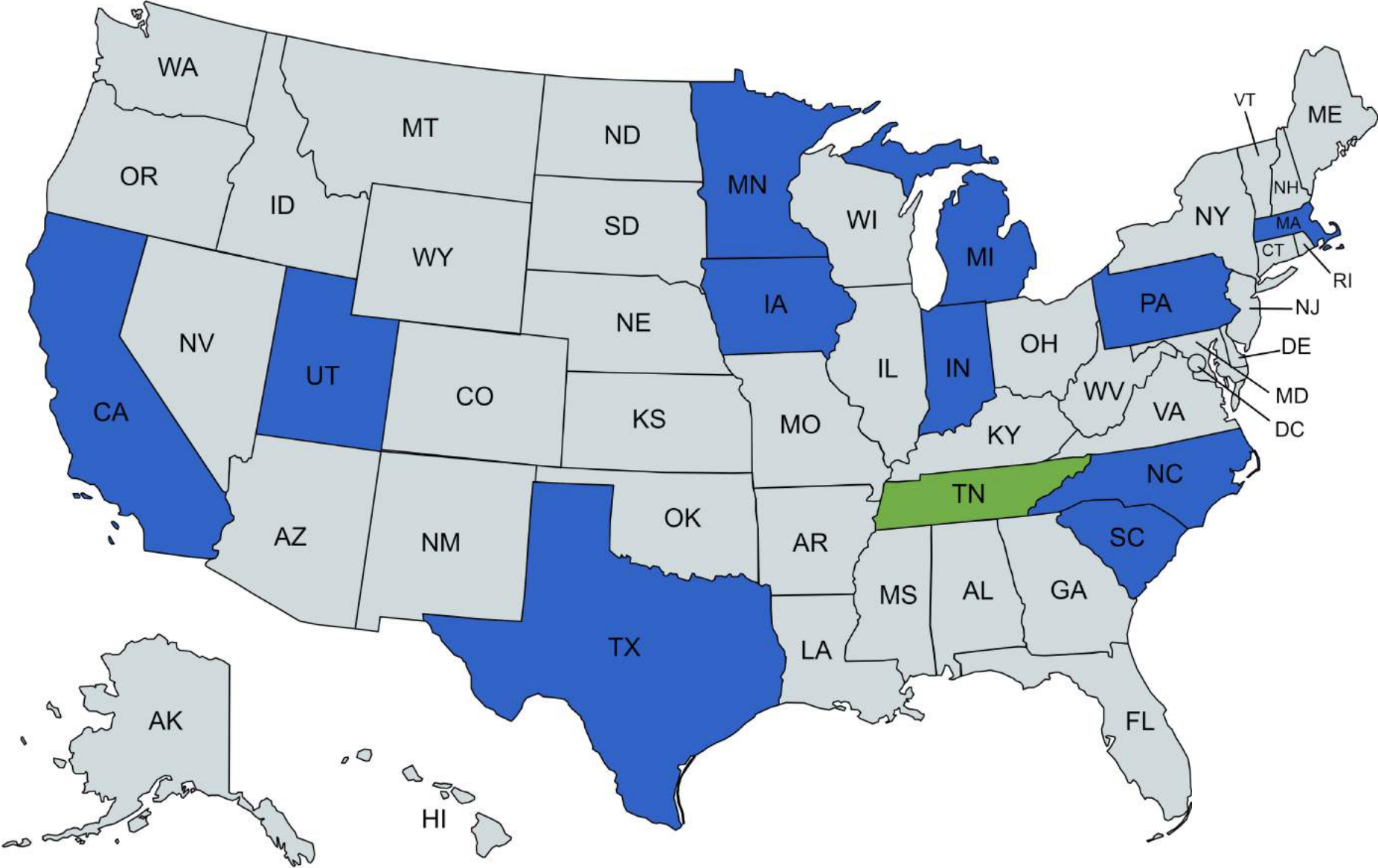
CustomerID	Year	Tickets for distinct events	Count of cities	Avg distance
494708	2022	5	10	6500 miles
	2023	6	12	



# IDENTIFYING POTENTIAL BROKERS



2023



## Inferences:

- Ticket could have been transferred or bought for somebody else
- Tickets could have been sold on the secondary market at higher prices

Championship Year: 2023  
CustomerID: 494708

■ Host cities  
■ Origin

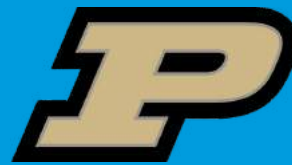


# PROBLEM STATEMENT 3



Where do we place the regional games in  
2027 and 2028?

# SUGGESTING LOCATIONS FOR REGIONALS 2027 AND 2028



## Selection criteria:

- Minimizing repeat selections of cities/states.
- Strong state interest in NCAA Div 1 women's basketball via Google Trends.
- Venues with at least 15,000 seats.
- Some consideration for state diversity.
- Regional venues must be neutral and not university arenas.

Suggestions	Seating	Women D1 BB Interest Rank	Last Hosted	Diversity Rank	Venue Type
Wells Fargo Arena, Iowa	16110	1	2008, 2012	47	Neutral
Capital One Arena, DC	20356	23	No Data	5	Neutral
PNC Arena, Charlotte, NC	19500	13	No Data	19	Neutral
Rupp Arena, Louisville, KY	23500	9	2017, 2018	43	Neutral



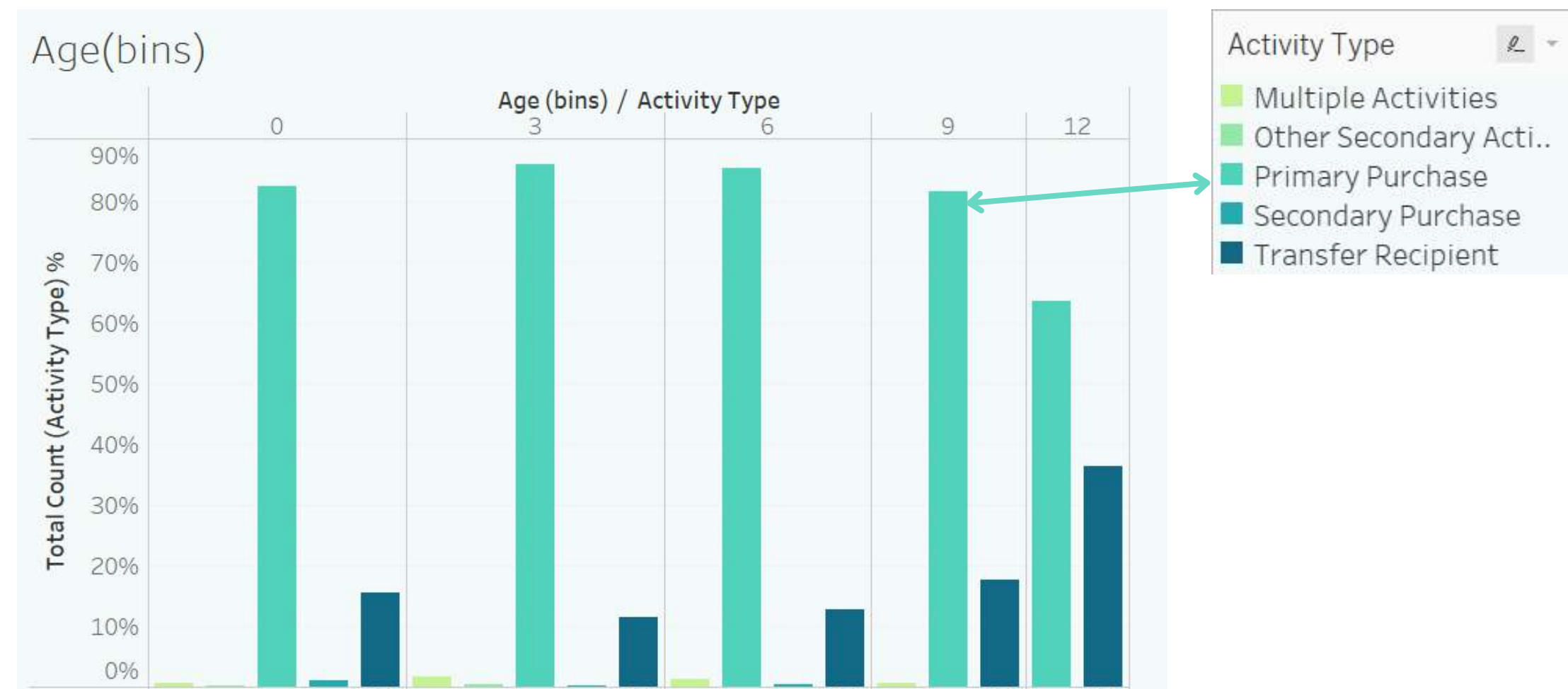
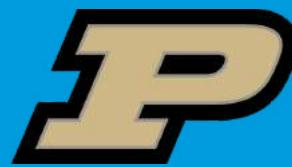
# PROBLEM STATEMENT 4



How do we identify who buys different types of tickets?



# TICKET ACTIVITY TYPES ACROSS SCENARIOS



- Primary Purchase is the highest bought activity across all age bins, with the highest coming from customers whose age on the platform is 4 or 10. (Age = Years since first interaction)

Event Round Name	Multiple Activities	Other Secondary Activity	Primary Purchase	Secondary Purchase	Transfer Recipient
Finals	4.0%	1.2%	27.1%	10.0%	57.7%
First and Second Rounds			100.0%		
Regionals	9.8%	1.2%	57.8%	9.7%	21.4%

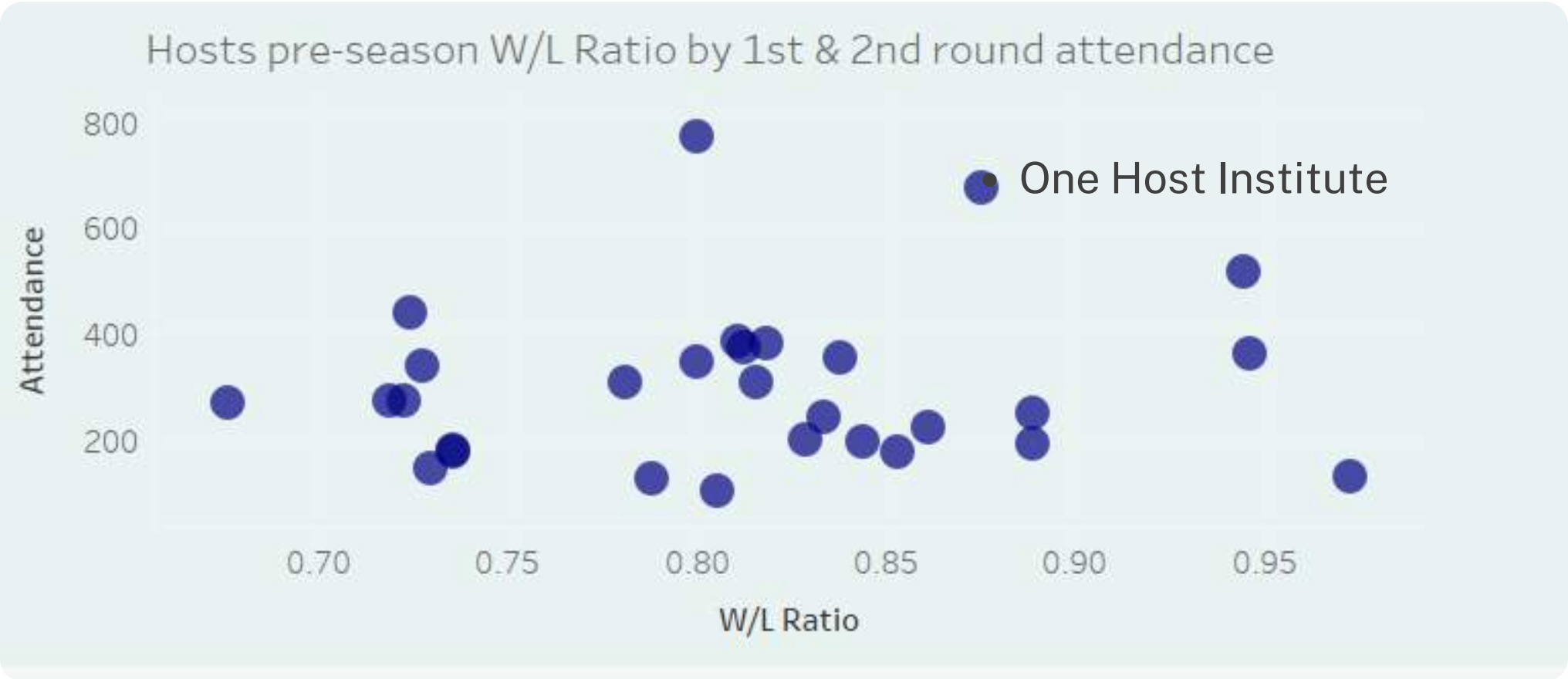
- % of tickets bought via Primary Purchase goes down as move from initial rounds to regionals and finals. This trend is something we intended our model to grasp.

# FACTORS INFLUENCING ATTENDANCE

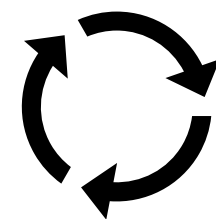
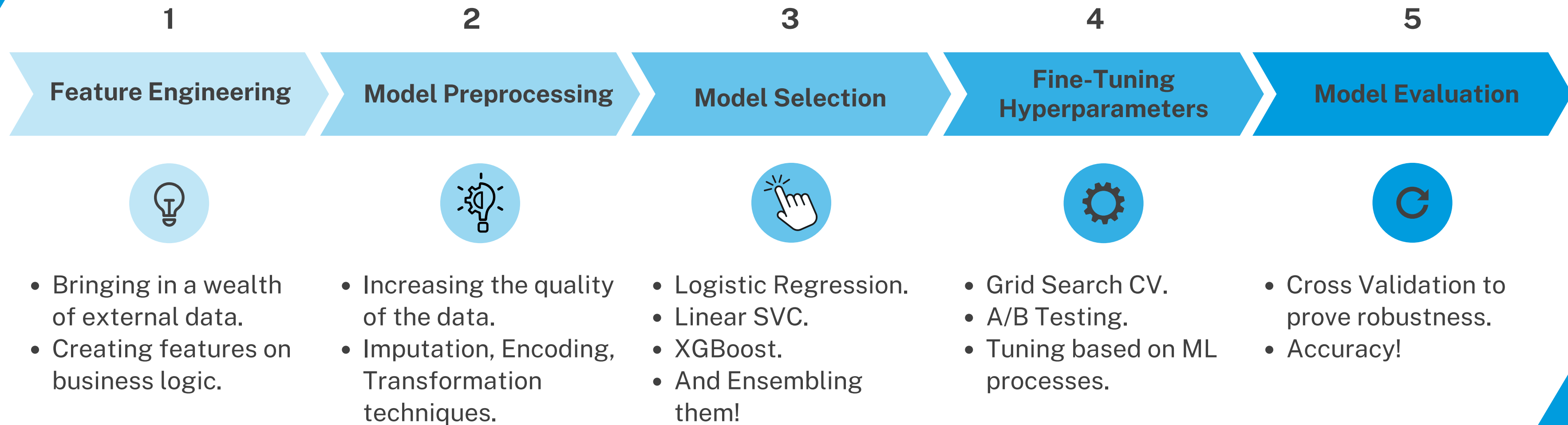


- Average attendance for the intial rounds is same across years while there is a surge in finals which could possibly signify the growth of the game.

- Previous W/L Ratio does not play a role in drawing the crowd.



# COMING TO OUR WINNING MODEL



**Used A/B testing to scientifically come up with better solutions**



# 1 Feature Engineering

External Sources: Usa.gov, NCAA.com, sports-reference.com, API's



## Distance

- Distance between the customer's location and the match location.



## Check on Zip Code and Teams Playing

- Check if the customer is present at the match location and hails from the same state as the competing teams.



## Conversion %

- Frequency of buying tickets(no of tickets bought/no of records in total).



## Age of customer

- Age of the customer on the platform since his 1st action.



## Prior Transactions

- Number of prior purchases just before the target transaction aggregated for that year.



## Average Gross Income per Annum

- Economic index of the origin location of the customer.

## 2 Model Preprocessing

### 2.DATA INCONSISTENCY CHECKS

- Inconsistent data types in rows converted.



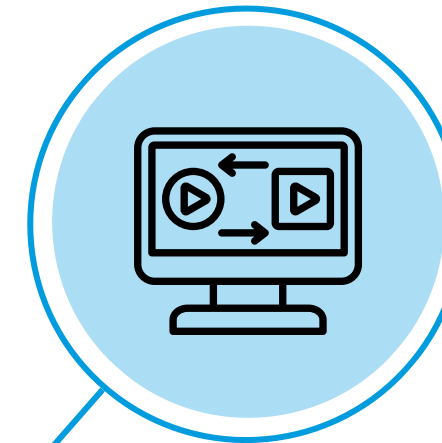
### 3.REMOVING REDUNDANT COLUMNS

- Columns like RecordID, Zipcodes etc.



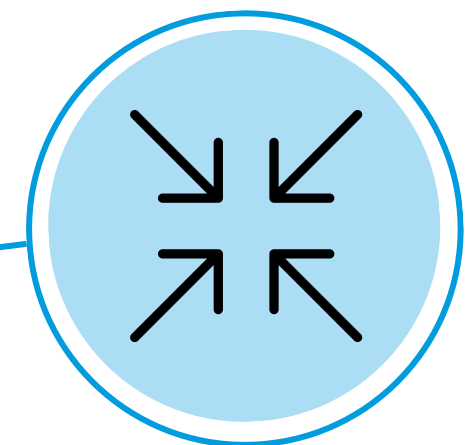
### 4.ENCODING

- Features are encoded into categories and labels



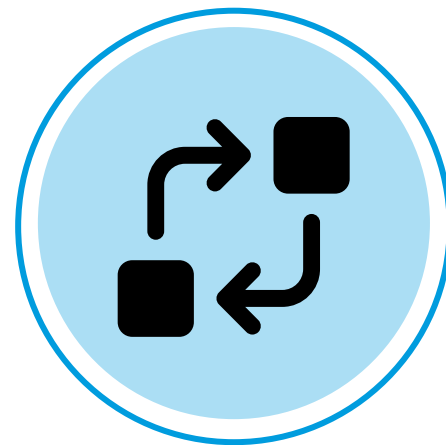
### 5.SCALING

- Applying standard scaling in order to standardize data



### 1.IMPUTING

- Replacing missing values by seasoned logic

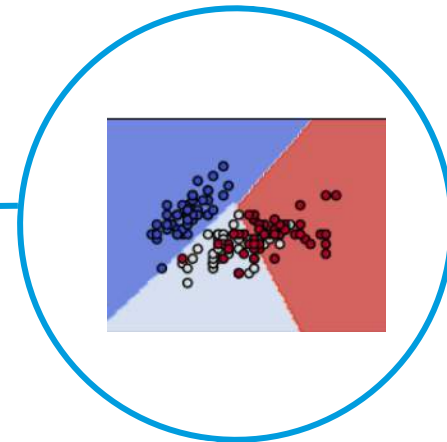






### Logistic Regression:

- Interpretable and efficient to train.
- Minimal computation



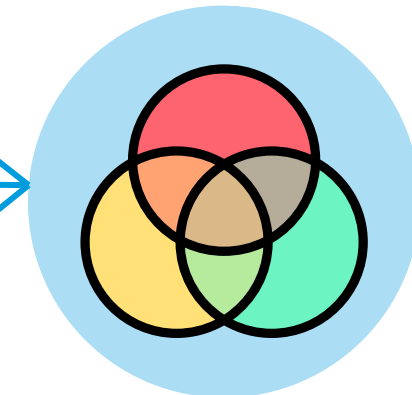
### Linear SVC:

- Helped us reduce dimensionality from encoded variables.
- Enhances generalization

**XG Boost**

### XG Boost

- Helps prevent overfitting.
- Can work on large datasets like our cleaned sparse training matrix.



### Ensembling Model

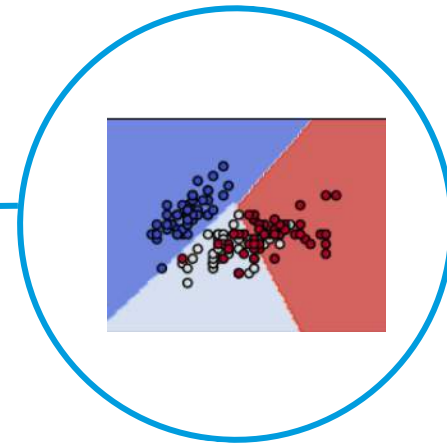
- Used ensembling with soft voting methods to get most likely predictions.
- Decreases the bias in outputs by considering multiple models.





### Logistic Regression:

- Max Iterations: 10,000
- Random State: 0
- C: 0.5



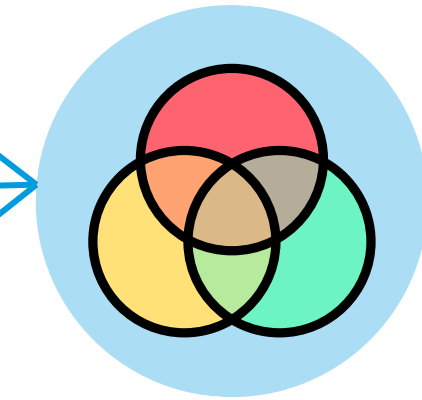
### Linear SVC:

- Wrapped in a calibrated SVC (CV=5)
- Random State: 96
- C: 1
- Tol(erance): 1e-3
- max\_iter: 15,000
- Dual: False



### XG Boost

- n\_estimators: 3000
- learning rate: 0.1
- Random State: 69



### Ensembling Model

- Voting: Soft. Used probability predictions to identify most probable class.

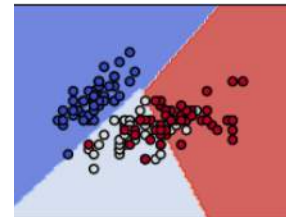
\*We also noticed accuracy scores higher on train and validation due to customer overlap not present in the test set.



Logistic  
Regression

Cross Validation  
Accuracy

**99.03%**



Linear  
SVC

Cross Validation  
Accuracy

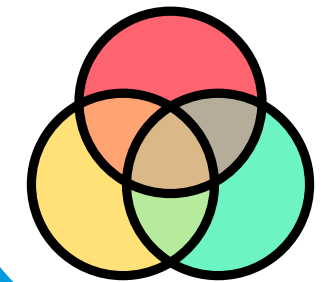
**99.10%**

*XG*  
*Boost*

XGBoost

Cross Validation  
Accuracy

**99.14%**



Ensemble






























Cross Validation  
score

**99.12%**











































High accuracy scores are due to the high number of 'No Activity' classes (about 91% of the data) that can be easily predicted. Actual accuracy lies around 86% for our best model



## Our Ranking - Public Leaderboard

#	Team	Members	Score	Entries	Last	Join
1	Green Light District	    	0.98795	59	16h	
 Your Best Entry! Your submission scored 0.98489, which is not an improvement of your previous score. Keep trying!						
2	First Up	  	0.98728	26	7h	
3	Pinnacles	    	0.98680	40	10h	
4	Purdue Champs	    	0.98671	48	9h	
5	Data Driven Developers	   	0.98661	44	10h	
6	Red Devils		0.98652	8	3d	
7	Precision Players	    	0.98642	27	11h	

## Our Ranking - Private Leaderboard

#	△	Team	Members	Score	Entries	Last	Solution
1	—	Green Light District	    	0.98682	59	6d	
2	—	First Up	    	0.98548	31	6d	
3	▲ 11	Datavizz	   	0.98520	4	10d	
4	▲ 16	BasketBots	    	0.98520	16	5d	
5	▼ 2	Pinnacles	    	0.98510	40	6d	
6	▲ 12	The Knowers	  	0.98501	27	6d	
7	▼ 3	Purdue Champs	    	0.98501	54	5d	
8	▲ 9	Kranniels School of Business	    	0.98481	30	5d	
9	▼ 1	Infosync	    	0.98472	36	5d	



# WHAT THE MODELS TELL US



What is helpful in predicting the ticket type?

MAJOR

Event Round Name is Missing\*

0.4090

Event is the final site

0.2712

Professional Sports Arenas

0.0741

OTHERS

Customer is not from the same city

0.0037

3rd week of March

0.0019

Purchase count\*

0.152



# WHAT WE RECOMMEND - PART 1



## Recommendation 1

- Soft Target - Leverage educational partnerships, local business collaborations, geo-targeted social media campaigns, and engagement with local sports clubs to boost NCAA women's basketball attendance.



## Recommendation 2

- Hard Target - Target those cities who are passionate about the game despite not hosting any games. NCAA could possibly host games in those locations.



# WHAT WE RECOMMEND - PART 2



## Recommendation 3

- Bookmark potential brokers and track their further activity to prevent them to try selling for profit.



## Recommendation 4

- Feedback Loops: Capture customer occupation, preferred communication methods, and referral sources to refine the model.







+ a b | e a u



# THANK YOU!



agarw402@purdue.edu  
nchidara@purdue.edu  
vponduri@purdue.edu  
sahoo14@purdue.edu  
dashd@purdue.edu



# APPENDIX





# APPENDIX 1: IMPORTANT FEATURES

## DISTINGUISHING TICKET TYPE

```
cat__EventRoundName_missing: 0.40902844071388245
cat__IsEventFinalSite_No: 0.2712523937225342
cat__FacilityDescription_Professional Sports Arena: 0.07418902963399887
cat__EventSession_All-Session: 0.030830474570393562
cat__EventRoundName_Finals: 0.008823391050100327
cat__EventRoundName Regionals: 0.004186241887509823
cat__Check_No: 0.0037762883584946394
num__ChampionshipYear: 0.003322502365335822
cat__EventRoundName_First and Second Rounds: 0.002695696661248803
cat__FacilityZipCode_missing: 0.0025133206509053707
cat__CustomerCity_missing: 0.0024835311342030764
cat__CustomerCity_Coralville: 0.002057366305962205
cat__startweek(m)_3: 0.001957931322976947
cat__CustomerFirstWBBActionDate_10-09-2020: 0.001554672489874065
cat__CustomerCity_Pittsburgh: 0.001552331494167447
cat__CustomerCity_Lakeway: 0.0014246187638491392
cat__CustomerCity_Reese: 0.0014176807599142194
cat__CustomerLastWBBActionDate_missing: 0.0013892744900658727
cat__CustomerCity_Elk Grove: 0.0013789274962618947
cat__CustomerCity_Tustin: 0.0013364702463150024
```

# APPENDIX 2: IMPORTANT FEATURES

## DISTINGUISHING PURCHASE/ NO PURCHASE

```
num__ChampionshipYear: 0.24709032475948334
cat__CustomerLastWBBActionDate_missing: 0.20466341078281403
num__purchase_count: 0.15293915569782257
num__Iowa st.: 0.04896533861756325
num__South Carolina: 0.042280636727809906
num__UNLV: 0.020647598430514336
cat__CustomerFirstWBBPurchaseDate_missing: 0.018807590007781982
num__Ohio st: 0.017612194642424583
num__FGCU: 0.01594054326415062
num__NC state: 0.014906602911651134
num__Maryland: 0.013601797632873058
num__Uconn: 0.013518132269382477
num__Georgia: 0.01308425609022379
num__Ole miss: 0.012940876185894012
num__North Carolina: 0.012828309088945389
num__Louisville: 0.01214776560664177
num__Stanford: 0.01167231984436512
num__Michigan: 0.01160008180886507
num__Texas: 0.010949882678687572
num__Indiana: 0.010930164717137814
num__Princeton: 0.010485007427632809
num__Arizona: 0.010322020389139652
num__Villanova: 0.009966501966118813
num__LSU: 0.00802601221948862
```



# APPENDIX 3: LINK TO TABLEAU DASHBOARD

[https://drive.google.com/drive/folders/  
1oGAboUNSa7nuwn0J7b49FuQhRKCC  
XLB5](https://drive.google.com/drive/folders/1oGAboUNSa7nuwn0J7b49FuQhRKCCXLB5)

# APPENDIX 4: LINK TO CODE AND MODEL

[https://drive.google.com/drive/folders/  
1pGwJKU7CDiZ3FSdwcrHWSIJdv89kF  
ThK?usp=sharing](https://drive.google.com/drive/folders/1pGwJKU7CDiZ3FSdwcrHWSIJdv89kFThK?usp=sharing)



# APPENDIX 5: POTENTIAL FRAUDULENT CUSTOMER 494708

CustomerID: 494708			
Year	Event dates	Distinct states tickest booked	Total tickets booked
2022	3/18/2022	5	5
	3/19/2022	2	2
	3/24/2022	1	1
	4/1/2022	1	1
	4/3/2022	1	1
2023	3/17/2023	7	8
	3/18/2023	1	1
	3/24/2023	1	3
	3/27/2023	1	1
	3/31/2023	1	1
	4/2/2023	1	1

# APPENDIX 6: GOOGLE TRENDS AND DIVERSITY INDEX

Sub Region	Trends Index	Diversity Index
Iowa	100	30.8
Connecticut	64	55.7
South Carolina	57	54.6
South Dakota	53	35.6
Indiana	36	41.3
Oregon	35	46.1
Mississippi	34	55.9
Vermont	32	20.2
Kentucky	31	32.8
Tennessee	31	46.6
Nebraska	31	40.8
Louisiana	29	58.6
North Carolina	27	57.9
Wyoming	26	32.4
Montana	25	30.1
Maryland	25	67.3
Kansas	24	45.4
North Dakota	24	32.6
Arkansas	23	49.8
Maine	23	18.5
Minnesota	23	40.5
Virginia	22	60.5
District of Columbia	21	67.2
Ohio	21	40.4
Delaware	21	59.6
Georgia	20	64.1
West Virginia	20	20.2

Sub Region	Trends Index	Diversity Index
Delaware	21	59.6
Georgia	20	64.1
West Virginia	20	20.2
Arizona	19	61.5
Oklahoma	18	59.5
Missouri	18	40.8
Rhode Island	18	49.4
Washington	18	55.9
Colorado	17	52.3
New Hampshire	17	23.6
New Mexico	17	63
Alabama	17	53.1
Massachusetts	16	51.6
Wisconsin	16	37
Illinois	16	60.3
Pennsylvania	16	44
Michigan	15	45.2
Idaho	15	35.9
Texas	15	67
Utah	15	40.7
Alaska	14	62.8
Florida	14	64.1
Nevada	14	68.8
New Jersey	14	65.8
Hawaii	14	76
New York	11	65.8
California	9	69.7