MACHINE LEARNING MADNESS

Project by

John Ellegard

NCAA MARCH MADNESS

- March Madness is one of the biggest, most exciting and most fun events in all of sports. The NCAA Division I men's basketball tournament is a single-elimination tournament of 68 teams that compete in seven rounds for the national championship.

- The first NCAA Division I men's basketball tournament was in 1939, and it has been held every year since.

- The inaugural tournament had just eight teams, and saw Oregon beat Ohio State 46-33 for the title. In 1951, the field doubled to 16, and kept expanding over the next few decades until 1985, when the modern format of a 64-team tournament began.

- March Madness was first used to refer to basketball by an Illinois high school official in 1939, but the term didn't find its way to the NCAA tournament until CBS broadcaster Brent Musburger used it during coverage of the 1982 tournament. The term has been synonymous with the NCAA Division I men's basketball tournament ever since.

MYBRACKET 2021

- The first NCAA bracket pool started in 1977 in a Staten Island bar. 88 people filled out brackets in the pool that year, and paid $10 in a winner-take-all format.
- At the same bar, in 2006, 150,000 entered, and prize money exceeded $1.5 million.
- In 2019, tens of millions of brackets were filled out through major online bracket games
- And every one of those millions of brackets has one goal - ***To be perfect….***
- How hard is it to pick a ***perfect*** bracket?

NCAA
.com

MYBRACKET 2021

A few more interesting facts about 9,223,372,036,854,775,808...

- There are 31.6 million seconds in a year, so 9.2 quintillion seconds is a quick 292 billion years.

- There have been 5 trillion days since the Big Bang, so repeat the entire history of our universe 1.8 million times.

- The Earth’s circumference is approximately 1.58 billion inches, so you’d have to walk around the planet 5.8 billion times.

- As of 2015, the best estimates for the number of trees on the planet was three trillion. Imagine that there was one single acorn hidden in one of those three trillion trees, and you were tasked with finding it on the first guess. Your odds of success are approximately three million times greater than picking a perfect bracket.

MYBRACKET 2021

DID YOU KNOW?

Famous investor Warren Buffett also got involved with the bracket game in 2014, when he joined forces with Quicken Loans to offer a $1 billion prize for anyone who picked a perfect bracket - of course - no one won.

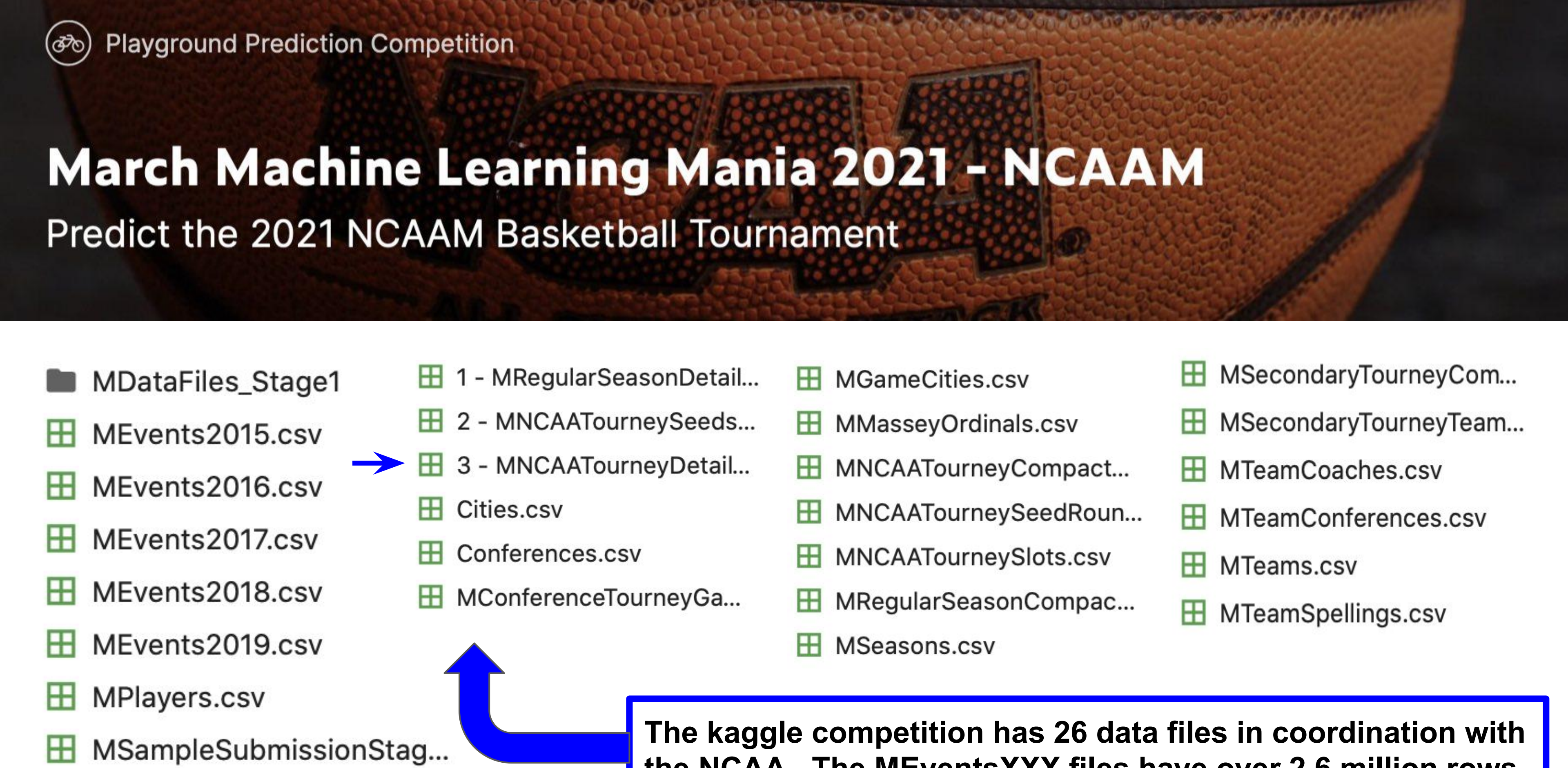
But in 2016, Buffett revived the contest only for employees of Berkshire Hathaway and its subsidiaries. The prize was also changed to $1 million every year for life to anyone who picked a perfect Sweet 16, which would be 48 correct picks in a row. No one claimed that one either.

Stage 1:

- Deadline: March 13th
- Train your model(s) with data from the regular season with data from the 1985 - 2014 regular seasons.
- Test your model(s) on the 2015-2019 NCAA Tournaments
- Allowed to use any external stats/resources you find useful in setting up your model(s),
- The trained model will be used in stage 2

Stage 2:

- Deadline: March 19th
- There are only a couple of days in Stage 2 in which to submit prediction files.
- The submission file for Stage 2 requires an entrant to forecast outcomes of all possible match-ups in the 2021 March Madness Tournament
- This is the final test in this competition to predict 2021 match results .



Playground Prediction Competition

March Machine Learning Mania 2021 - NCAAM

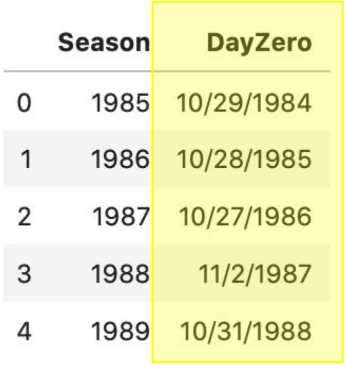
Predict the 2021 NCAAM Basketball Tournament

- MDataFiles_Stage1
- MEvents2015.csv
- MEvents2016.csv
- MEvents2017.csv
- MEvents2018.csv
- MEvents2019.csv
- MPlayers.csv
- MSampleSubmissionStag...
- 1 - MRegularSeasonDetail...
- 2 - MNCAATourneySeeds...
- 3 - MNCAATourneyDetail...
- Cities.csv
- Conferences.csv
- MConferenceTourneyGa...
- MGameCities.csv
- MMasseyOrdinals.csv
- MNCAATourneyCompact...
- MNCAATourneySeedRoun...
- MNCAATourneySlots.csv
- MRegularSeasonCompac...
- MSeasons.csv
- MSecondaryTourneyCom...
- MSecondaryTourneyTeam...
- MTeamCoaches.csv
- MTeamConferences.csv
- MTeams.csv
- MTeamSpellings.csv

The kaggle competition has 26 data files in coordination with the NCAA. The MEventsXXX files have over 2.6 million rows.

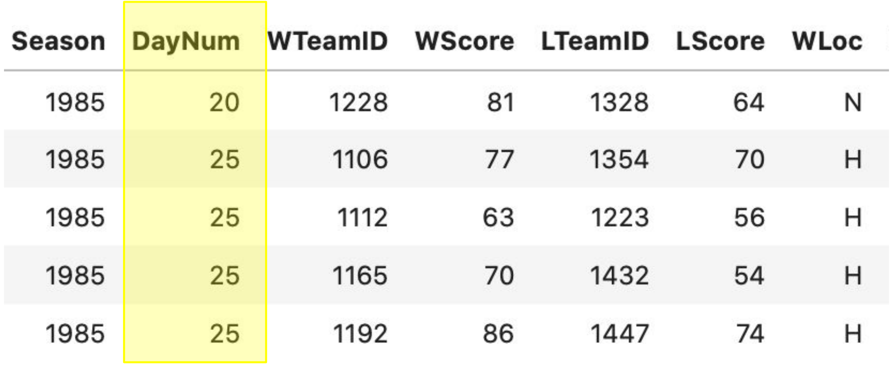
NCAA

Google Cloud



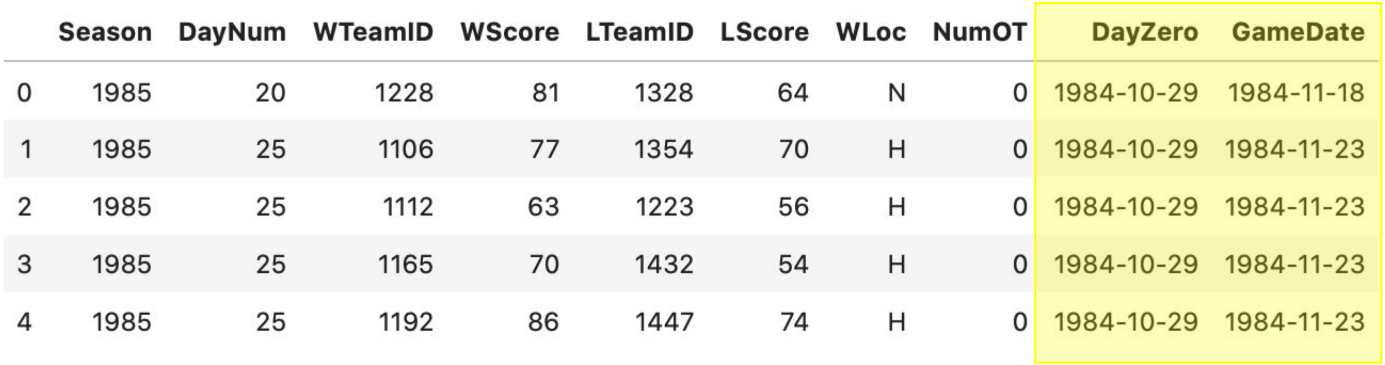
	Season	DayZero
0	1985	10/29/1984
1	1986	10/28/1985
2	1987	10/27/1986
3	1988	11/2/1987
4	1989	10/31/1988

+



Season	DayNum	WTeamID	WScore	LTeamID	LScore	WLoc
1985	20	1228	81	1328	64	N
1985	25	1106	77	1354	70	H
1985	25	1112	63	1223	56	H
1985	25	1165	70	1432	54	H
1985	25	1192	86	1447	74	H

=

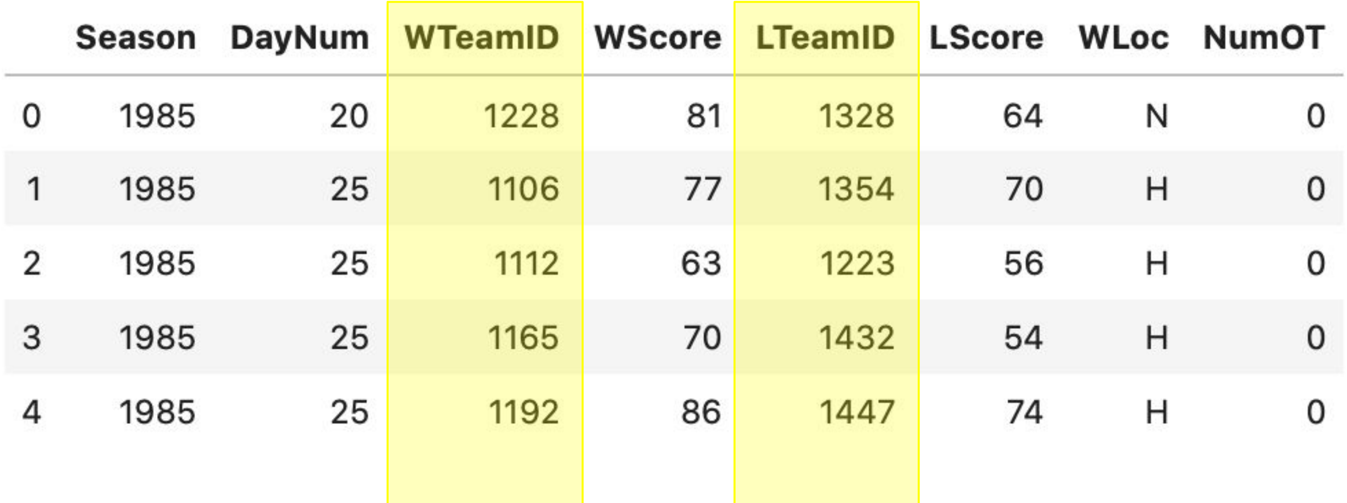


	Season	DayNum	WTeamID	WScore	LTeamID	LScore	WLoc	NumOT	DayZero	GameDate
0	1985	20	1228	81	1328	64	N	0	1984-10-29	1984-11-18
1	1985	25	1106	77	1354	70	H	0	1984-10-29	1984-11-23
2	1985	25	1112	63	1223	56	H	0	1984-10-29	1984-11-23
3	1985	25	1165	70	1432	54	H	0	1984-10-29	1984-11-23
4	1985	25	1192	86	1447	74	H	0	1984-10-29	1984-11-23

The NCAA bases their schedule on the DayZero column. DayZero tells you the date corresponding to DayNum = 0 during that season. All game dates are aligned upon a common scale so that each year:

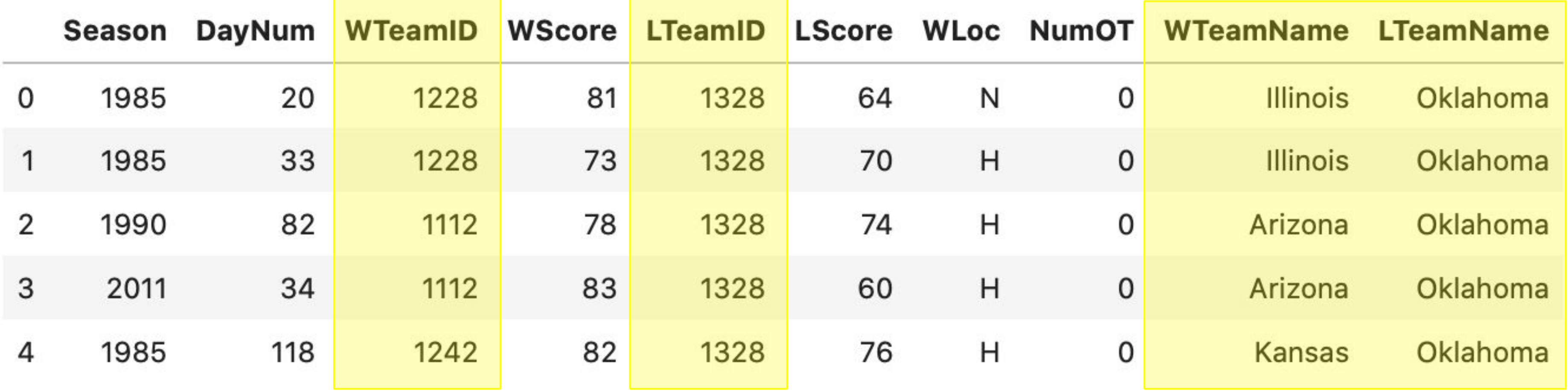
- DayNum = 132 is Selection Sunday
- DayNum = 132 is also the final day of the regular season
- DayNum = 134/135 are the days they "play-in" games are played
- DayNum = 152 is the day the National Semifinals are always on
- DayNum = 154 is the Monday of the Championship Game

MRegularSeasonCompactResults.csv

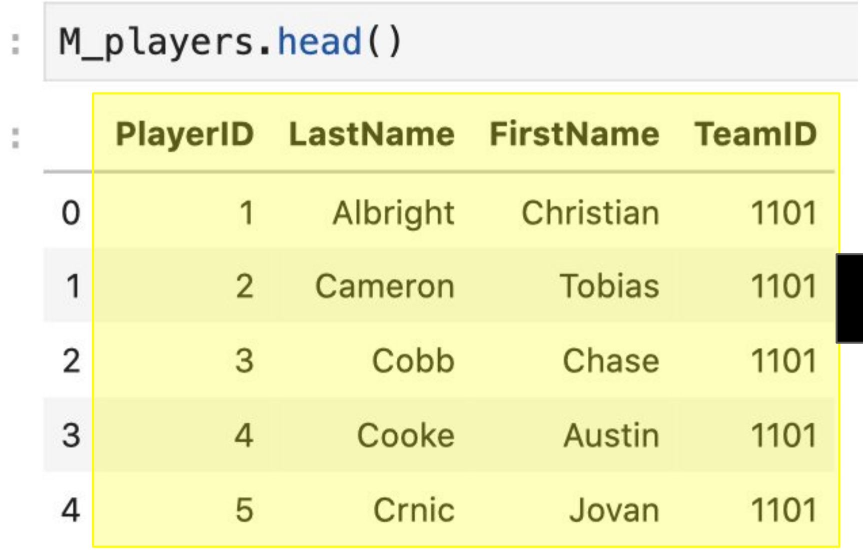


	Season	DayNum	WTeamID	WScore	LTeamID	LScore	WLoc	NumOT
0	1985	20	1228	81	1328	64	N	0
1	1985	25	1106	77	1354	70	H	0
2	1985	25	1112	63	1223	56	H	0
3	1985	25	1165	70	1432	54	H	0
4	1985	25	1192	86	1447	74	H	0

With team names from MTeams.csv merged...



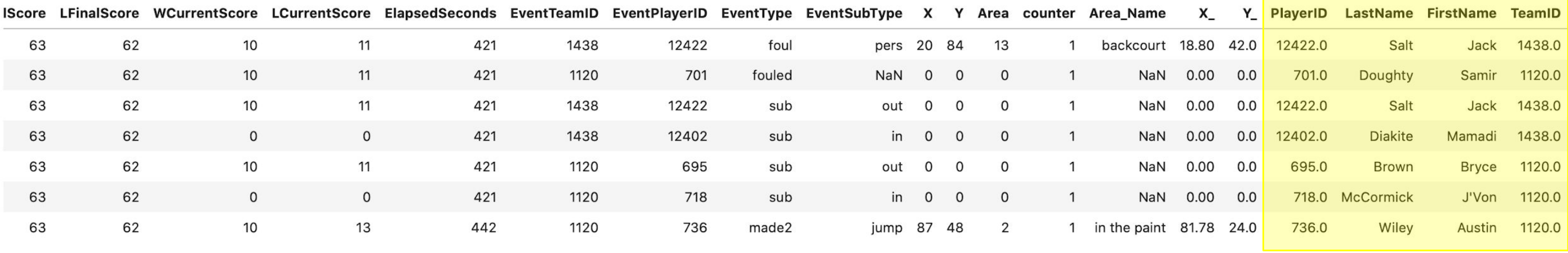
	Season	DayNum	WTeamID	WScore	LTeamID	LScore	WLoc	NumOT	WTeamName	LTeamName
0	1985	20	1228	81	1328	64	N	0	Illinois	Oklahoma
1	1985	33	1228	73	1328	70	H	0	Illinois	Oklahoma
2	1990	82	1112	78	1328	74	H	0	Arizona	Oklahoma
3	2011	34	1112	83	1328	60	H	0	Arizona	Oklahoma
4	1985	118	1242	82	1328	76	H	0	Kansas	Oklahoma



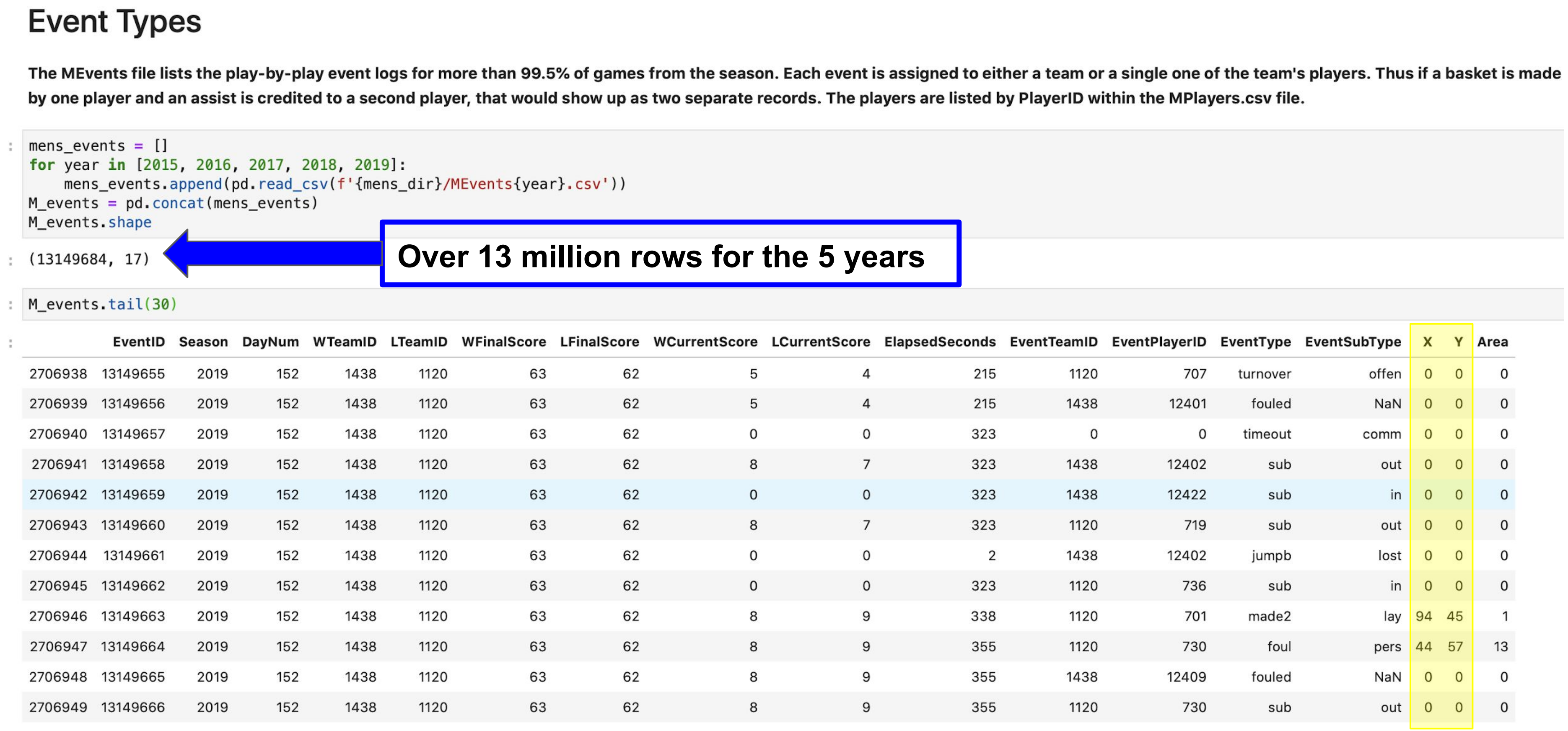
```
M_players.head()
```

	PlayerID	LastName	FirstName	TeamID
0	1	Albright	Christian	1101
1	2	Cameron	Tobias	1101
2	3	Cobb	Chase	1101
3	4	Cooke	Austin	1101
4	5	Crnic	Jovan	1101

Merge "M_players" onto "M_events" so we can look up any player's events by player's name versus "PlayerID".



lScore	LFinalScore	WCurrentScore	LCurrentScore	ElapsedSeconds	EventTeamID	EventPlayerID	EventType	EventSubType	X	Y	Area	counter	Area_Name	X_	Y_	PlayerID	LastName	FirstName	TeamID
63	62	10	11	421	1438	12422	foul	pers	20	84	13	1	backcourt	18.80	42.0	12422.0	Salt	Jack	1438.0
63	62	10	11	421	1120	701	fouled	NaN	0	0	0	1	NaN	0.00	0.0	701.0	Doughty	Samir	1120.0
63	62	10	11	421	1438	12422	sub	out	0	0	0	1	NaN	0.00	0.0	12422.0	Salt	Jack	1438.0
63	62	0	0	421	1438	12402	sub	in	0	0	0	1	NaN	0.00	0.0	12402.0	Diakite	Mamadi	1438.0
63	62	10	11	421	1120	695	sub	out	0	0	0	1	NaN	0.00	0.0	695.0	Brown	Bryce	1120.0
63	62	0	0	421	1120	718	sub	in	0	0	0	1	NaN	0.00	0.0	718.0	McCormick	J'Von	1120.0
63	62	10	13	442	1120	736	made2	jump	87	48	2	1	in the paint	81.78	24.0	736.0	Wiley	Austin	1120.0



Event Types

The MEvents file lists the play-by-play event logs for more than 99.5% of games from the season. Each event is assigned to either a team or a single one of the team's players. Thus if a basket is made by one player and an assist is credited to a second player, that would show up as two separate records. The players are listed by PlayerID within the MPlayers.csv file.

```
mens_events = []
for year in [2015, 2016, 2017, 2018, 2019]:
    mens_events.append(pd.read_csv(f'{mens_dir}/MEvents{year}.csv'))
M_events = pd.concat(mens_events)
M_events.shape
```

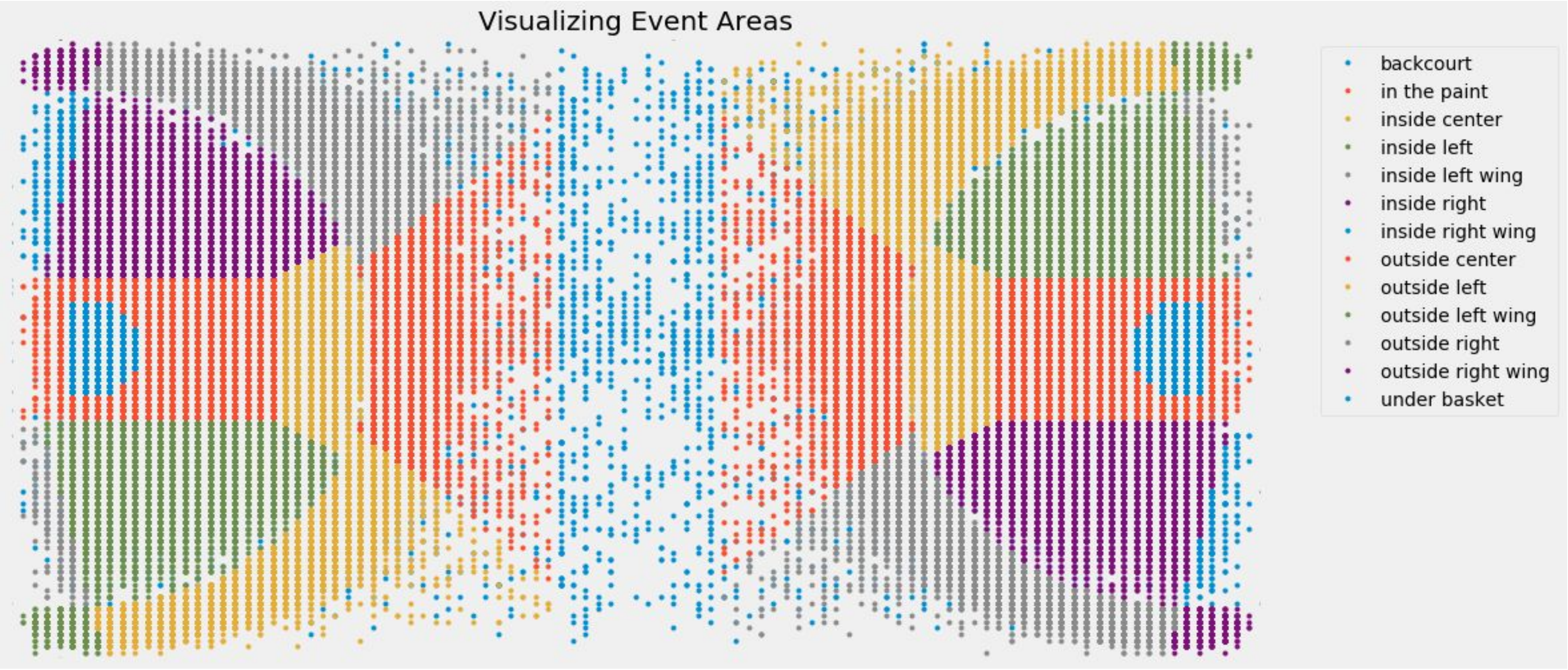
```
(13149684, 17)
```

Over 13 million rows for the 5 years

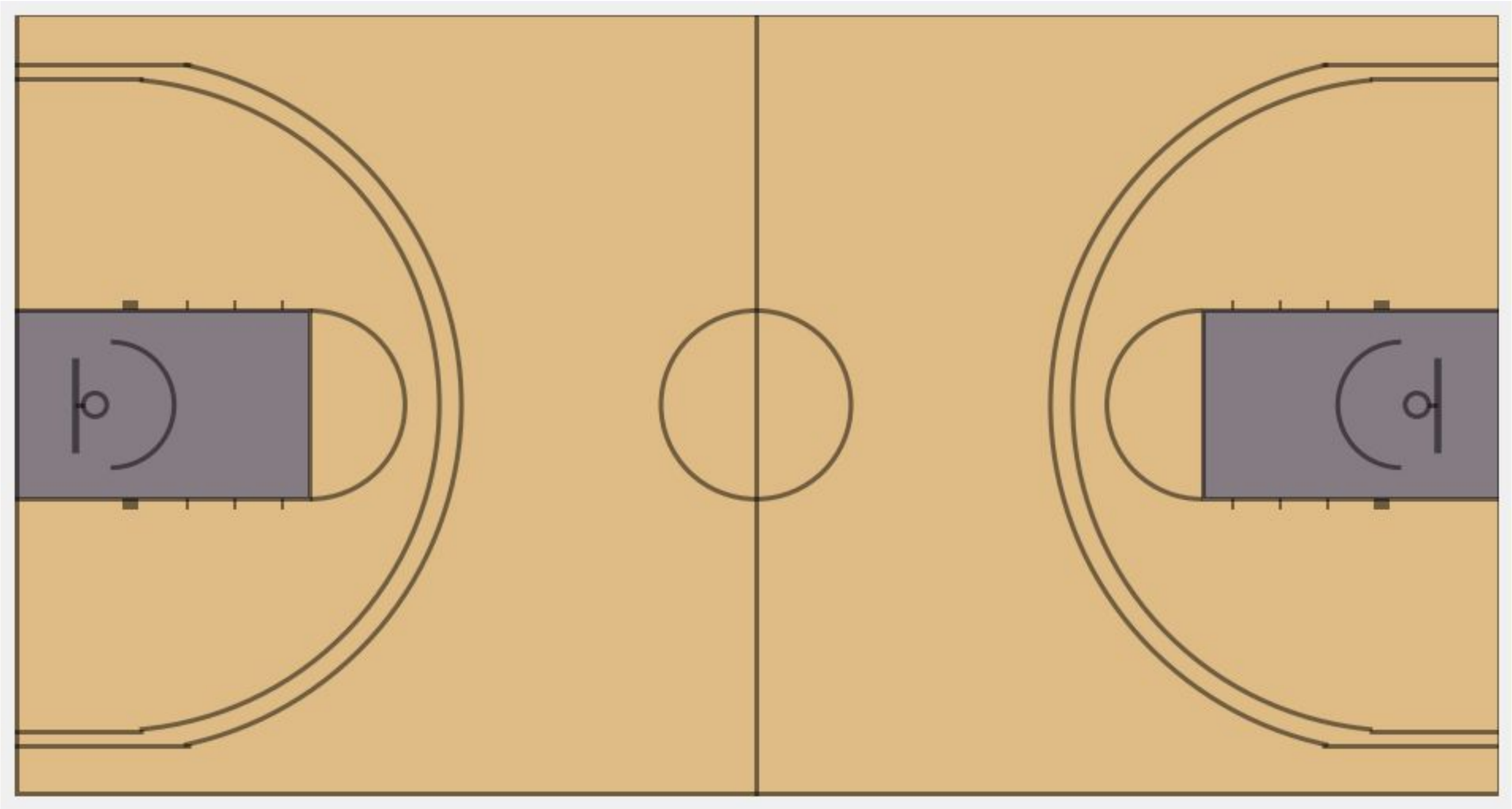
```
M_events.tail(30)
```

	EventID	Season	DayNum	WTeamID	LTeamID	WFinalScore	LFinalScore	WCurrentScore	LCurrentScore	ElapsedSeconds	EventTeamID	EventPlayerID	EventType	EventSubType	X	Y	Area
2706938	13149655	2019	152	1438	1120	63	62	5	4	215	1120	707	turnover	offen	0	0	0
2706939	13149656	2019	152	1438	1120	63	62	5	4	215	1438	12401	fouled	NaN	0	0	0
2706940	13149657	2019	152	1438	1120	63	62	0	0	323	0	0	timeout	comm	0	0	0
2706941	13149658	2019	152	1438	1120	63	62	8	7	323	1438	12402	sub	out	0	0	0
2706942	13149659	2019	152	1438	1120	63	62	0	0	323	1438	12422	sub	in	0	0	0
2706943	13149660	2019	152	1438	1120	63	62	8	7	323	1120	719	sub	out	0	0	0
2706944	13149661	2019	152	1438	1120	63	62	0	0	2	1438	12402	jumpb	lost	0	0	0
2706945	13149662	2019	152	1438	1120	63	62	0	0	323	1120	736	sub	in	0	0	0
2706946	13149663	2019	152	1438	1120	63	62	8	9	338	1120	701	made2	lay	94	45	1
2706947	13149664	2019	152	1438	1120	63	62	8	9	355	1120	730	foul	pers	44	57	13
2706948	13149665	2019	152	1438	1120	63	62	8	9	355	1438	12409	fouled	NaN	0	0	0
2706949	13149666	2019	152	1438	1120	63	62	8	9	355	1120	730	sub	out	0	0	0

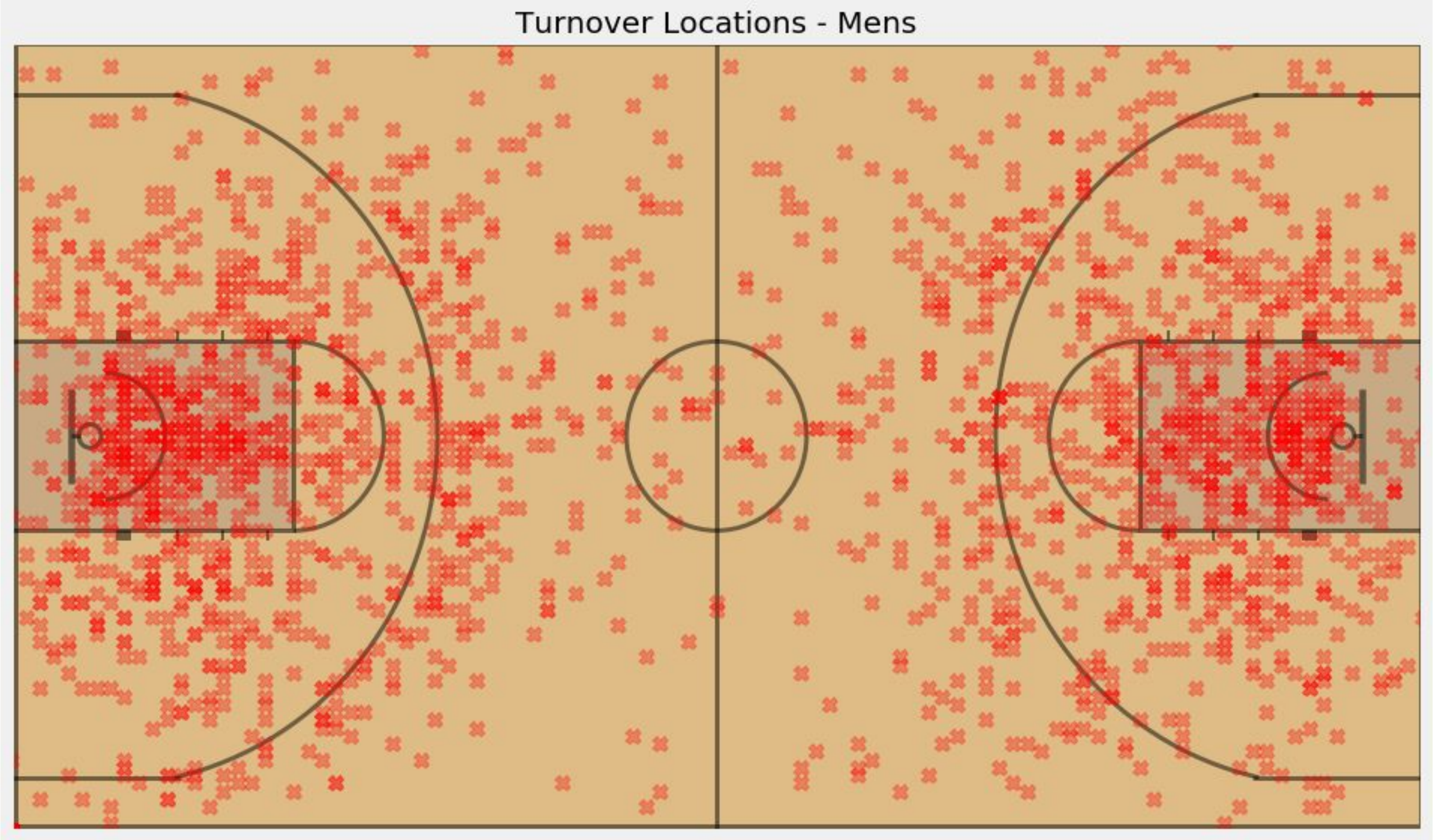
With the events file we can create some really cool visuals. This visual just creates a map of events with no concern as to what the event is.

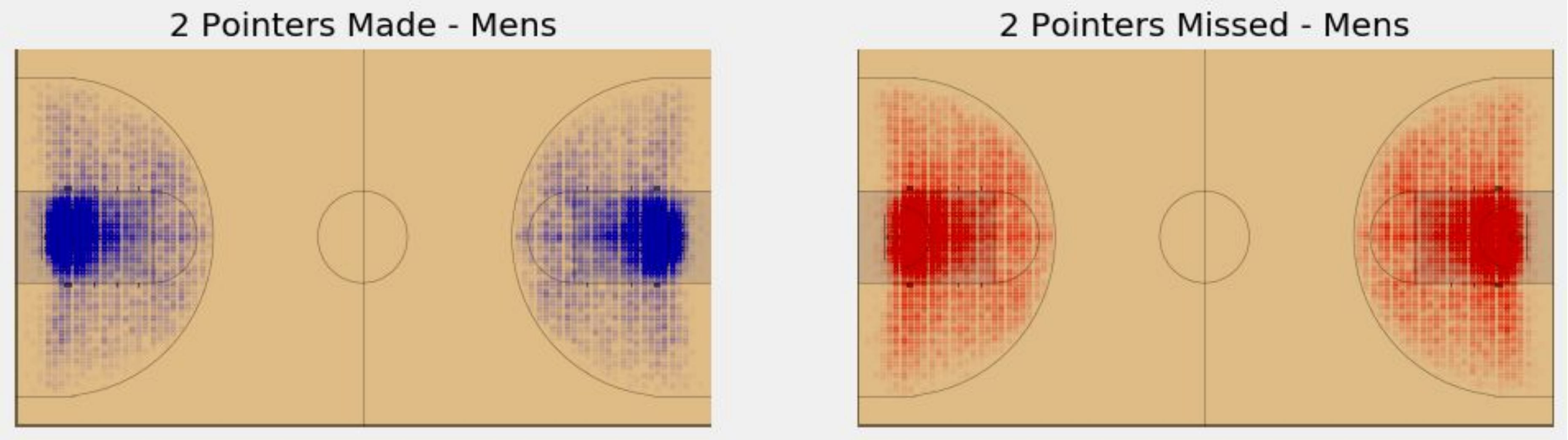


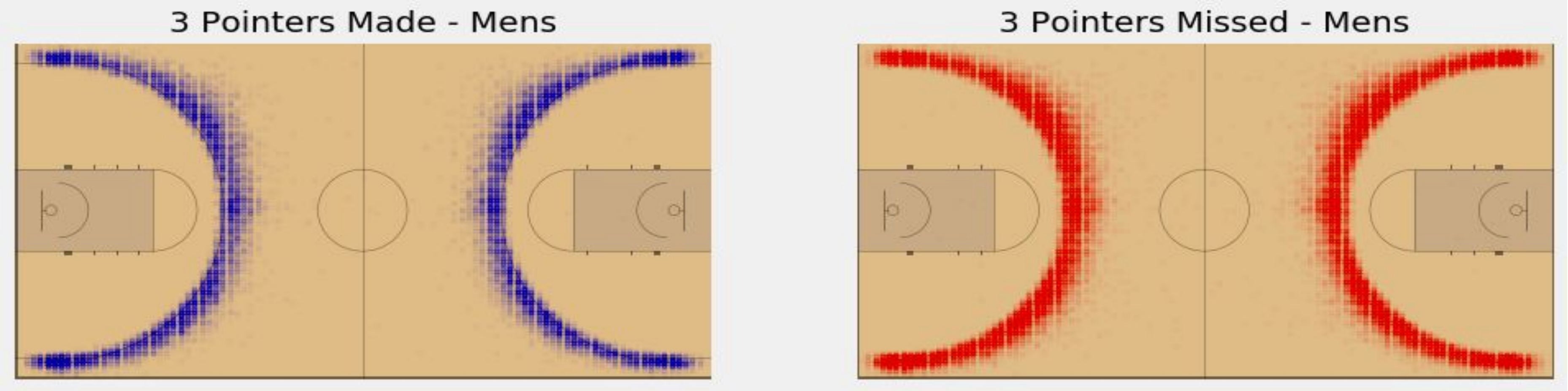
This blank court is generated from code created by Rob Mulla.
https://github.com/RobMulla

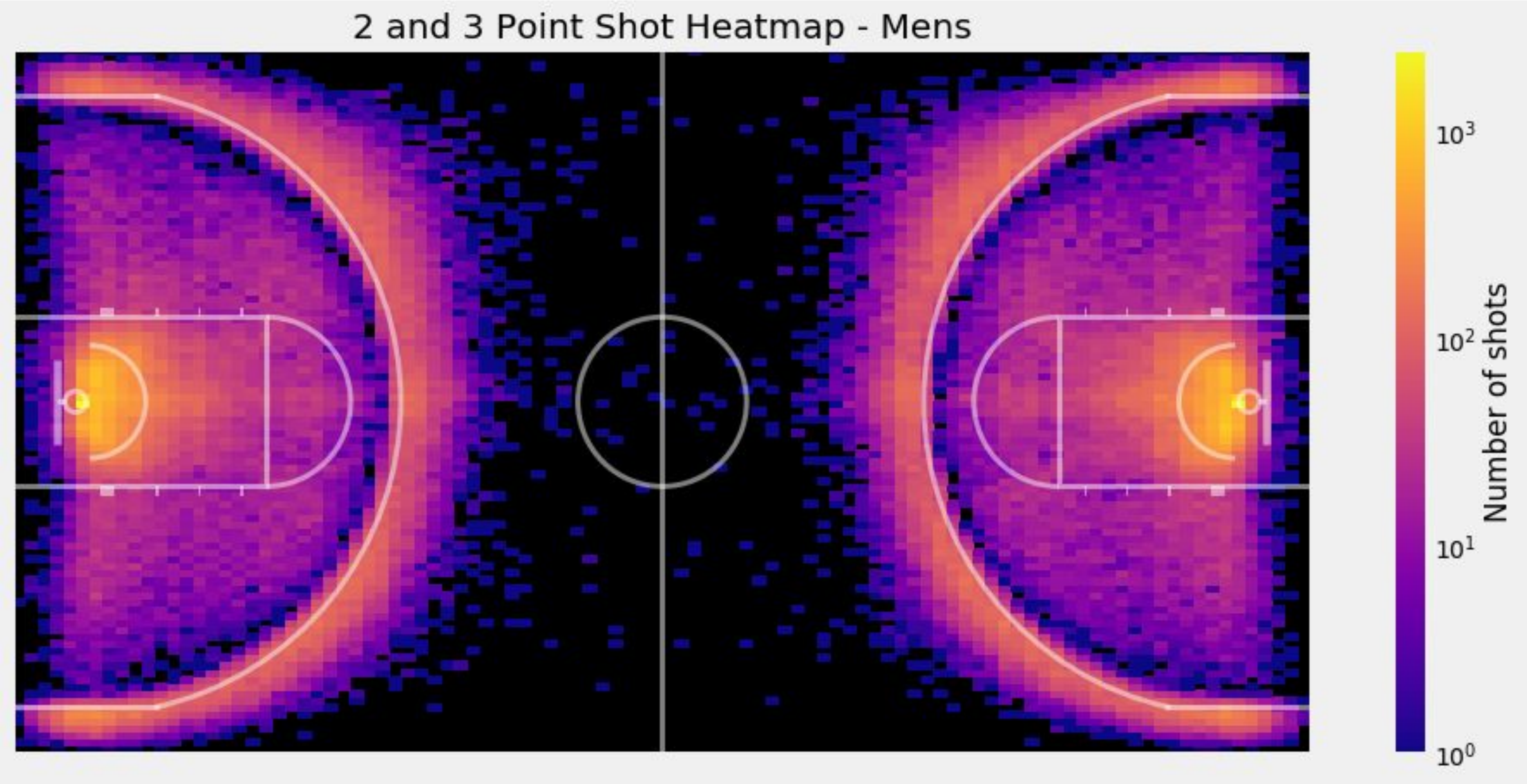


With the blank court and the events file we can create some interesting visualizations...

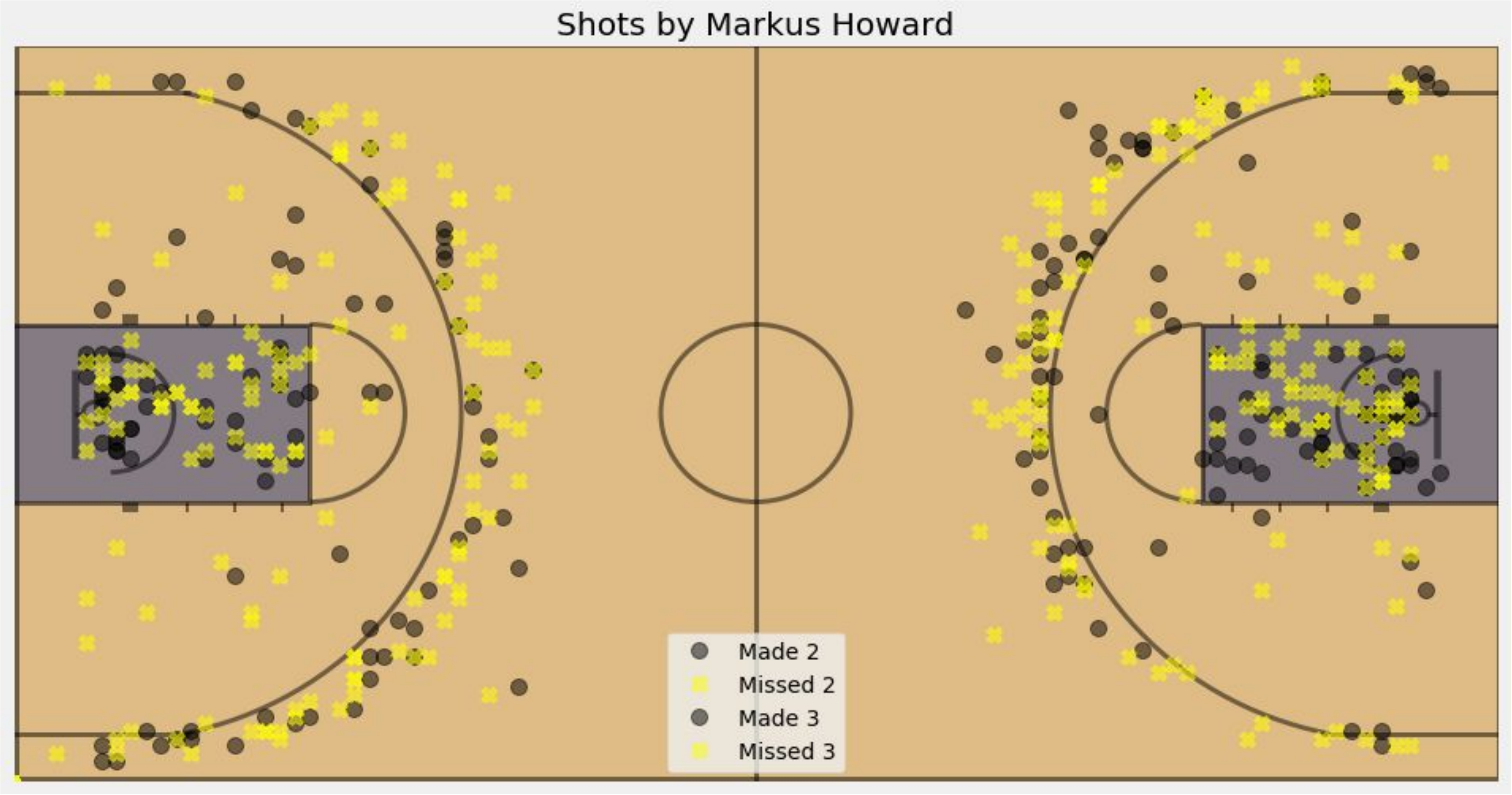



2 Pointers Made - Mens
2 Pointers Missed - Mens

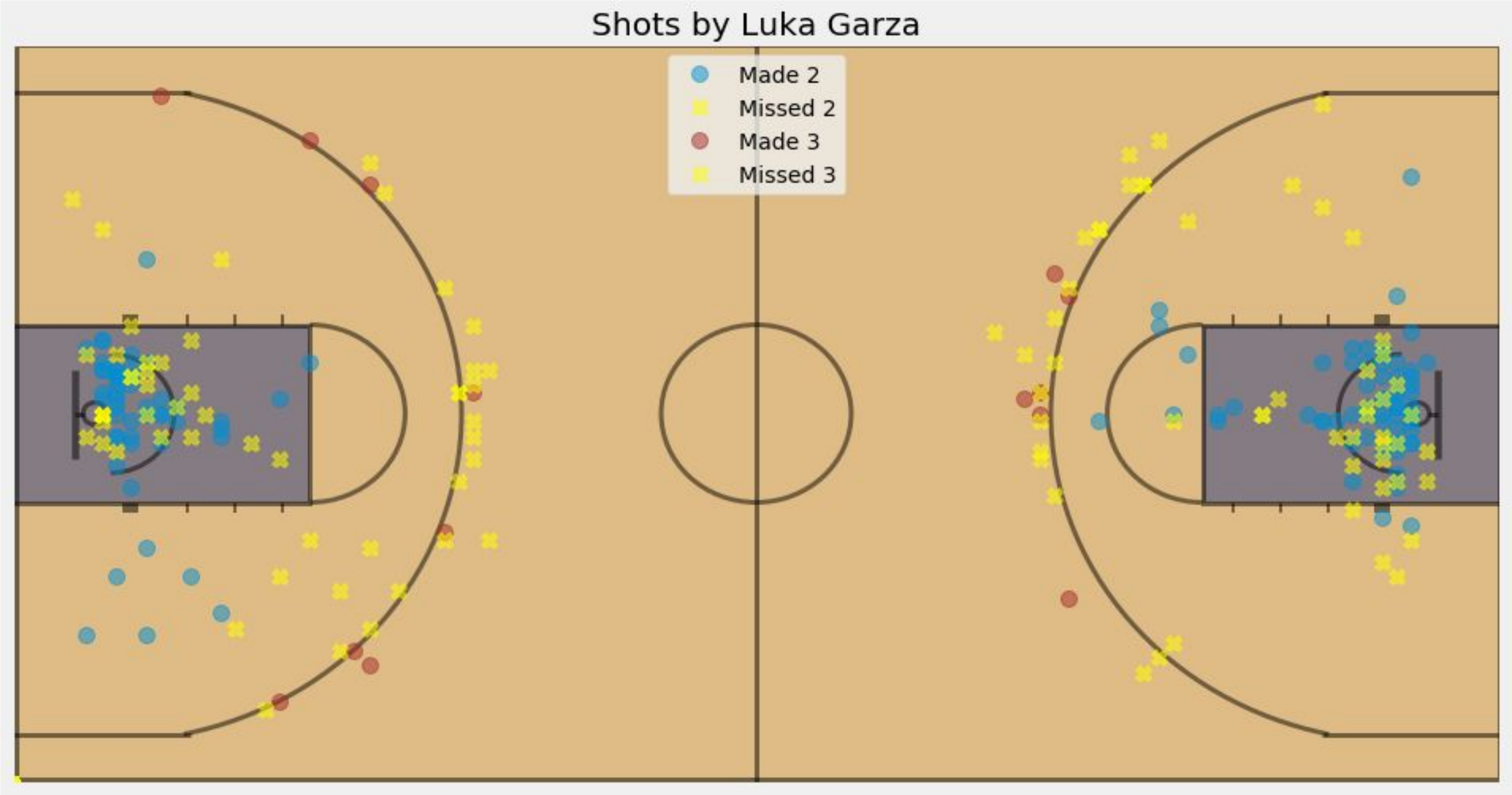

3 Pointers Made - Mens
3 Pointers Missed - Mens


2 and 3 Point Shot Heatmap - Mens
Number of shots
10^3
10^2
10^1
10^0

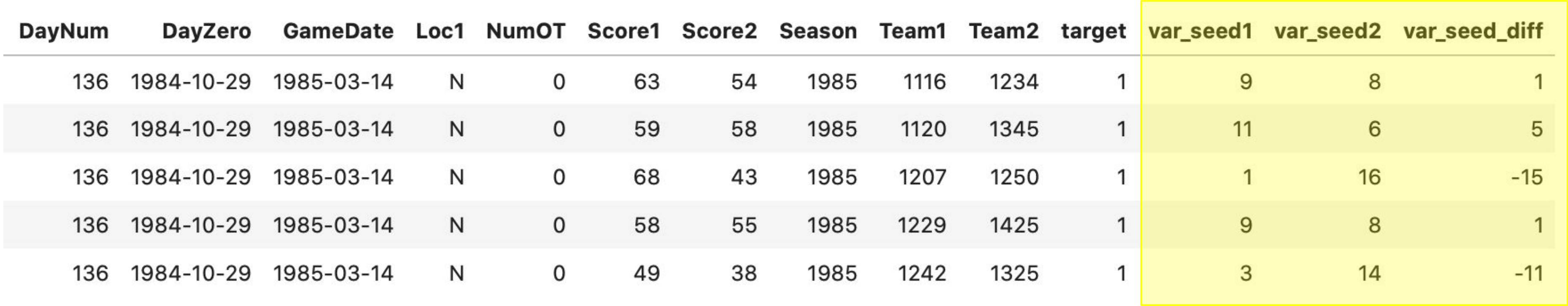
Combining the blank court with data from Markus Howard



Combining the blank court with data from Luka Garza



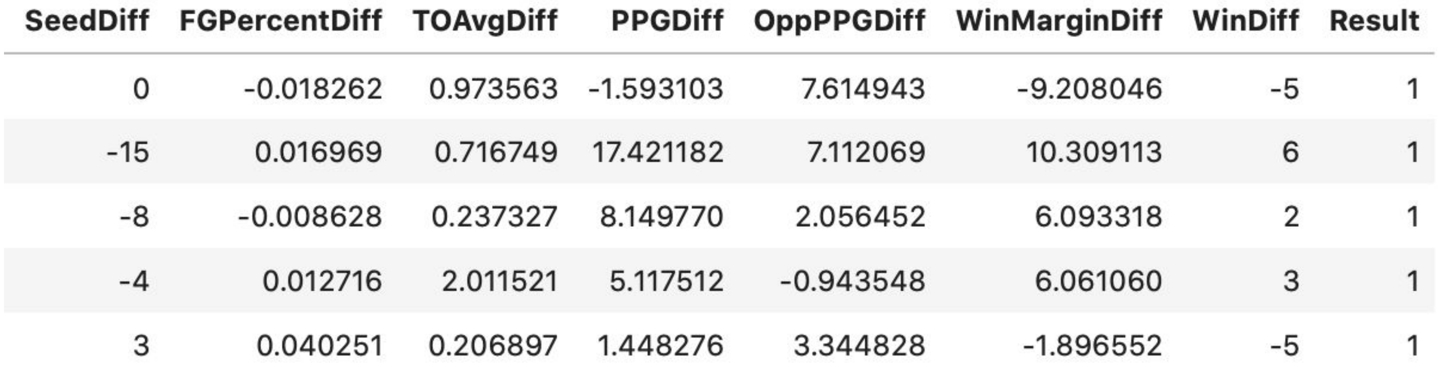
Logistic Regression



DayNum	DayZero	GameDate	Loc1	NumOT	Score1	Score2	Season	Team1	Team2	target	var_seed1	var_seed2	var_seed_diff
136	1984-10-29	1985-03-14	N	0	63	54	1985	1116	1234	1	9	8	1
136	1984-10-29	1985-03-14	N	0	59	58	1985	1120	1345	1	11	6	5
136	1984-10-29	1985-03-14	N	0	68	43	1985	1207	1250	1	1	16	-15
136	1984-10-29	1985-03-14	N	0	58	55	1985	1229	1425	1	9	8	1
136	1984-10-29	1985-03-14	N	0	49	38	1985	1242	1325	1	3	14	-11

log loss: -0.535411

Grid Search with Multiple Models



SeedDiff	FGPercentDiff	TOAvgDiff	PPGDiff	OppPPGDiff	WinMarginDiff	WinDiff	Result
0	-0.018262	0.973563	-1.593103	7.614943	-9.208046	-5	1
-15	0.016969	0.716749	17.421182	7.112069	10.309113	6	1
-8	-0.008628	0.237327	8.149770	2.056452	6.093318	2	1
-4	0.012716	2.011521	5.117512	-0.943548	6.061060	3	1
3	0.040251	0.206897	1.448276	3.344828	-1.896552	-5	1

Gradient Boosting Classifier: -0.5601

K-Nearest Neighbors Classifier: -0.5617

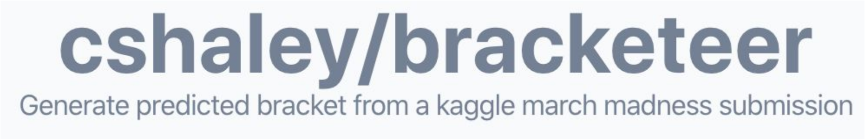
Random Forest Classifier: -0.5951

Support Vector Classification: -0.5484

Logistic Regression: -0.5486

NCAA®

Google Cloud



cshaley/bracketeer

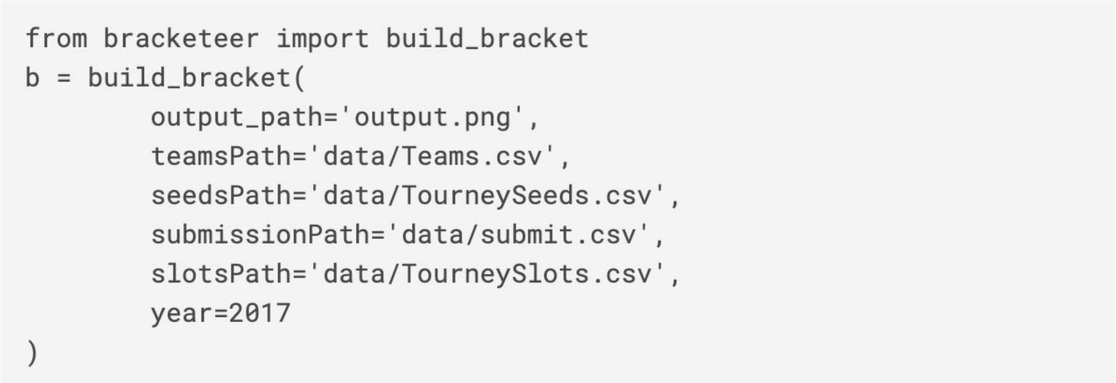
Generate predicted bracket from a kaggle march madness submission

Charlie Haley

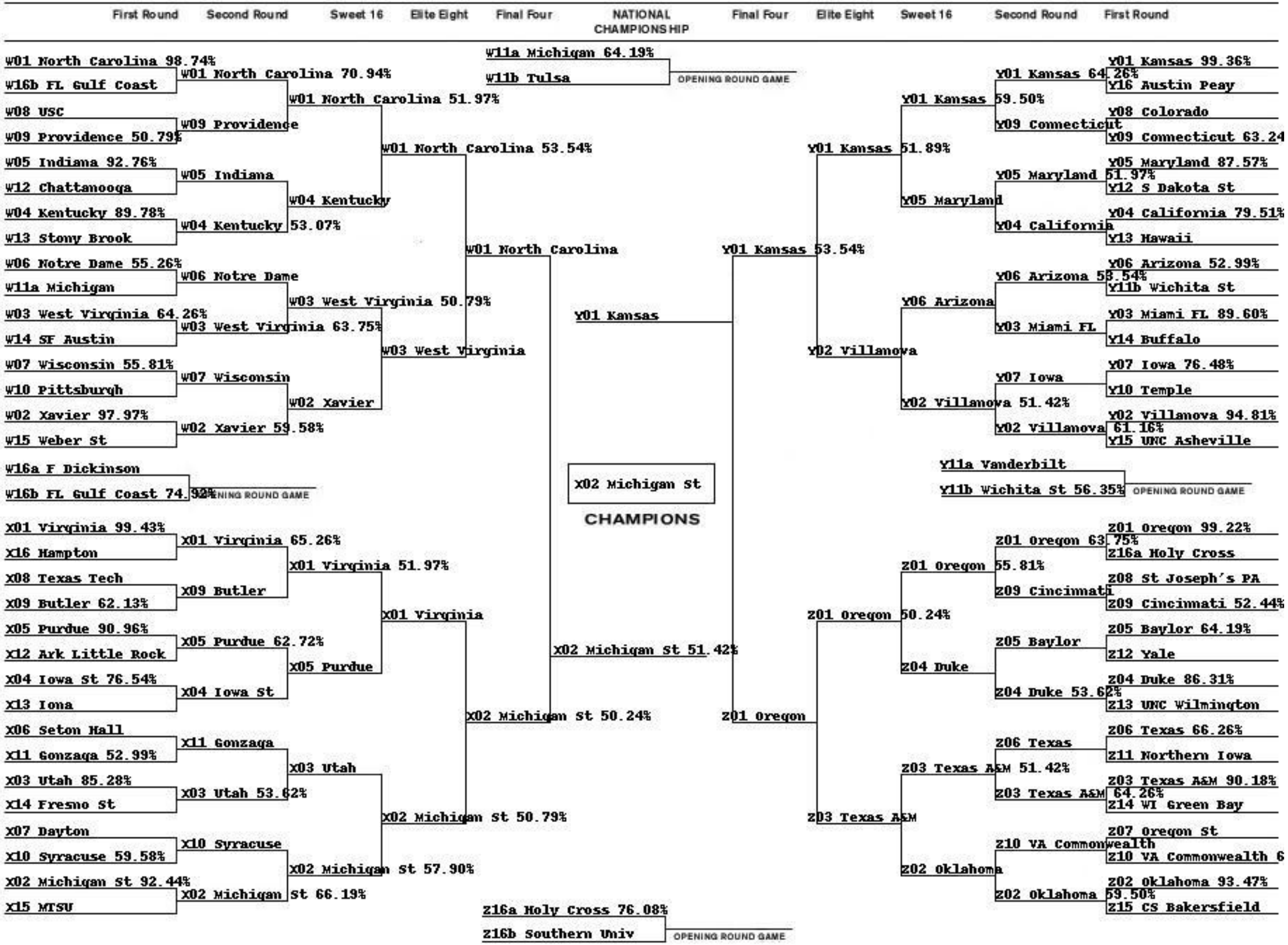
cshaley

Dallas, TX

https://www.linkedin.com/in/charliehaley



```
from bracketeer import build_bracket
b = build_bracket(
        output_path='output.png',
        teamsPath='data/Teams.csv',
        seedsPath='data/TourneySeeds.csv',
        submissionPath='data/submit.csv',
        slotsPath='data/TourneySlots.csv',
        year=2017
)
```



First Round
Second Round
Sweet 16
Elite Eight
Final Four
NATIONAL CHAMPIONSHIP
Final Four
Elite Eight
Sweet 16
Second Round
First Round

W11a Michigan 64.19%
W11b Tulsa
OPENING ROUND GAME

W01 North Carolina 98.74%
W16b FL Gulf Coast
W08 USC
W09 Providence 50.79%
W05 Indiana 92.76%
W12 Chattanooga
W04 Kentucky 89.78%
W13 Stony Brook
W06 Notre Dame 55.26%
W11a Michigan
W03 West Virginia 64.26%
W14 SF Austin
W07 Wisconsin 55.81%
W10 Pittsburgh
W02 Xavier 97.97%
W15 Weber St

W01 North Carolina 70.94%
W09 Providence
W05 Indiana
W04 Kentucky 53.07%
W06 Notre Dame
W03 West Virginia 63.75%
W07 Wisconsin
W02 Xavier 59.58%

W01 North Carolina 51.97%
W04 Kentucky
W03 West Virginia 50.79%
W02 Xavier

W01 North Carolina 53.54%
W03 West Virginia

W01 North Carolina

W16a F Dickinson
W16b FL Gulf Coast 74.92%
OPENING ROUND GAME

X01 Virginia 99.43%
X16 Hampton
X08 Texas Tech
X09 Butler 62.13%
X05 Purdue 90.96%
X12 Ark Little Rock
X04 Iowa St 76.54%
X13 Iona
X06 Seton Hall
X11 Gonzaga 52.99%
X03 Utah 85.28%
X14 Fresno St
X07 Dayton
X10 Syracuse 59.58%
X02 Michigan St 92.44%
X15 MTSU

X01 Virginia 65.26%
X09 Butler
X05 Purdue 62.72%
X04 Iowa St
X11 Gonzaga
X03 Utah 53.62%
X10 Syracuse
X02 Michigan St 66.19%

X01 Virginia 51.97%
X05 Purdue
X03 Utah
X02 Michigan St 57.90%

X01 Virginia
X02 Michigan St 50.79%

X02 Michigan St 50.24%

Y01 Kansas
X02 Michigan St 51.42%

X02 Michigan St

CHAMPIONS

Y01 Kansas 53.54%
Z01 Oregon

Y01 Kansas 51.89%
Y02 Villanova

Y01 Kansas 59.50%
Y05 Maryland
Y06 Arizona
Y02 Villanova 51.42%

Y01 Kansas 64.26%
Y09 Connecticut
Y05 Maryland 51.97%
Y04 California
Y06 Arizona 53.54%
Y03 Miami FL
Y07 Iowa
Y02 Villanova 61.16%

Y01 Kansas 99.36%
Y16 Austin Peay
Y08 Colorado
Y09 Connecticut 63.24
Y05 Maryland 87.57%
Y12 S Dakota St
Y04 California 79.51%
Y13 Hawaii
Y06 Arizona 52.99%
Y11b Wichita St
Y03 Miami FL 89.60%
Y14 Buffalo
Y07 Iowa 76.48%
Y10 Temple
Y02 Villanova 94.81%
Y15 UNC Asheville

Y11a Vanderbilt
Y11b Wichita St 56.35%
OPENING ROUND GAME

Z01 Oregon 50.24%
Z03 Texas A&M

Z01 Oregon 55.81%
Z04 Duke
Z03 Texas A&M 51.42%
Z02 Oklahoma

Z01 Oregon 63.75%
Z09 Cincinnati
Z05 Baylor
Z04 Duke 53.62%
Z06 Texas
Z03 Texas A&M 64.26%
Z10 VA Commonwealth
Z02 Oklahoma 59.50%

Z01 Oregon 99.22%
Z16a Holy Cross
Z08 St Joseph's PA
Z09 Cincinnati 52.44%
Z05 Baylor 64.19%
Z12 Yale
Z04 Duke 86.31%
Z13 UNC Wilmington
Z06 Texas 66.26%
Z11 Northern Iowa
Z03 Texas A&M 90.18%
Z14 WI Green Bay
Z07 Oregon St
Z10 VA Commonwealth 6
Z02 Oklahoma 93.47%
Z15 CS Bakersfield

Z16a Holy Cross 76.08%
Z16b Southern Univ
OPENING ROUND GAME

KEEP CALM

AND

LET THE

MADNESS

BEGIN