Codedex Hackathon

Track 3: Predict 2024 Olympics Champion

Nancy Shih

Codédex

PARIS 2024

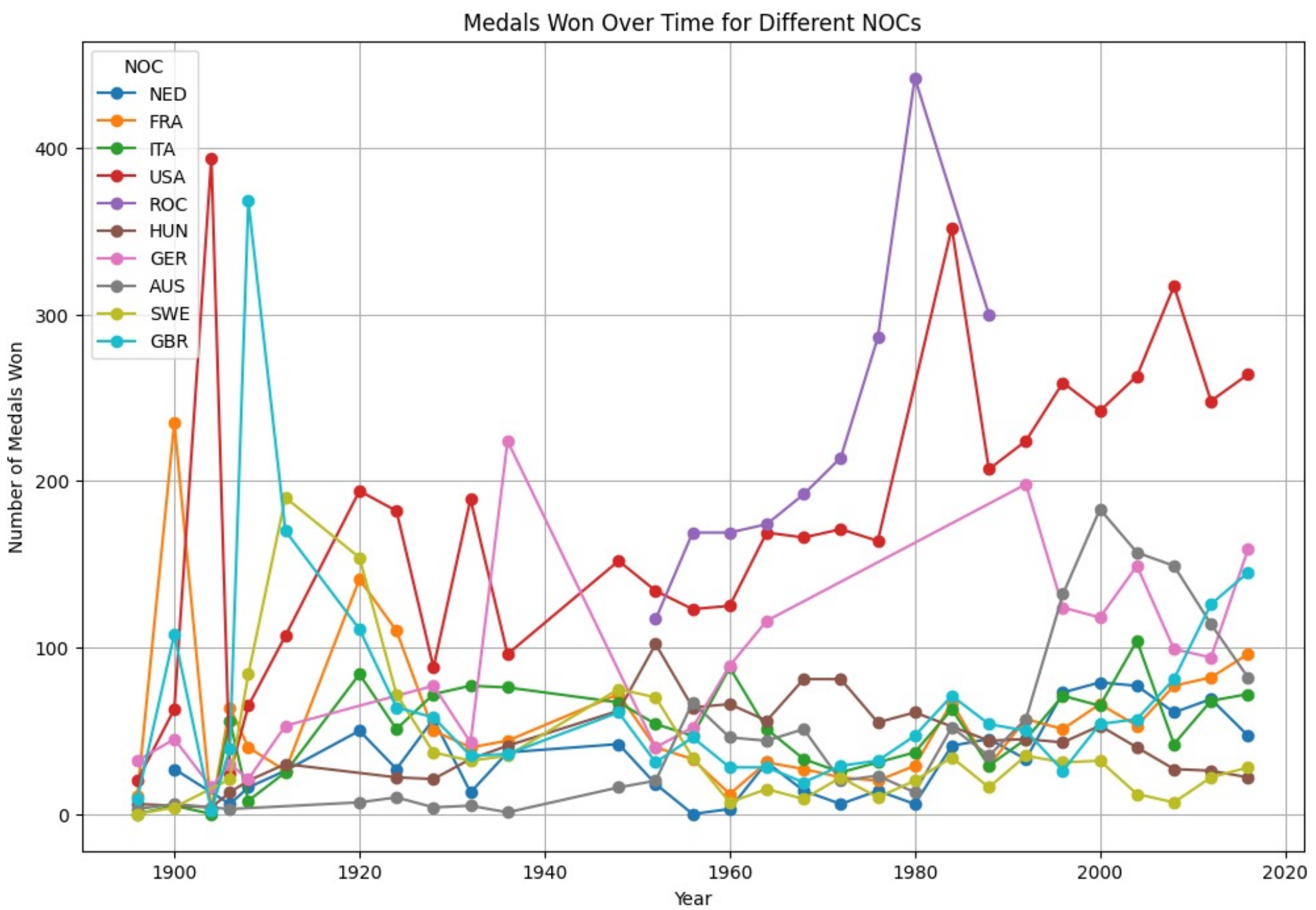
120 years Olympics Data Analytics

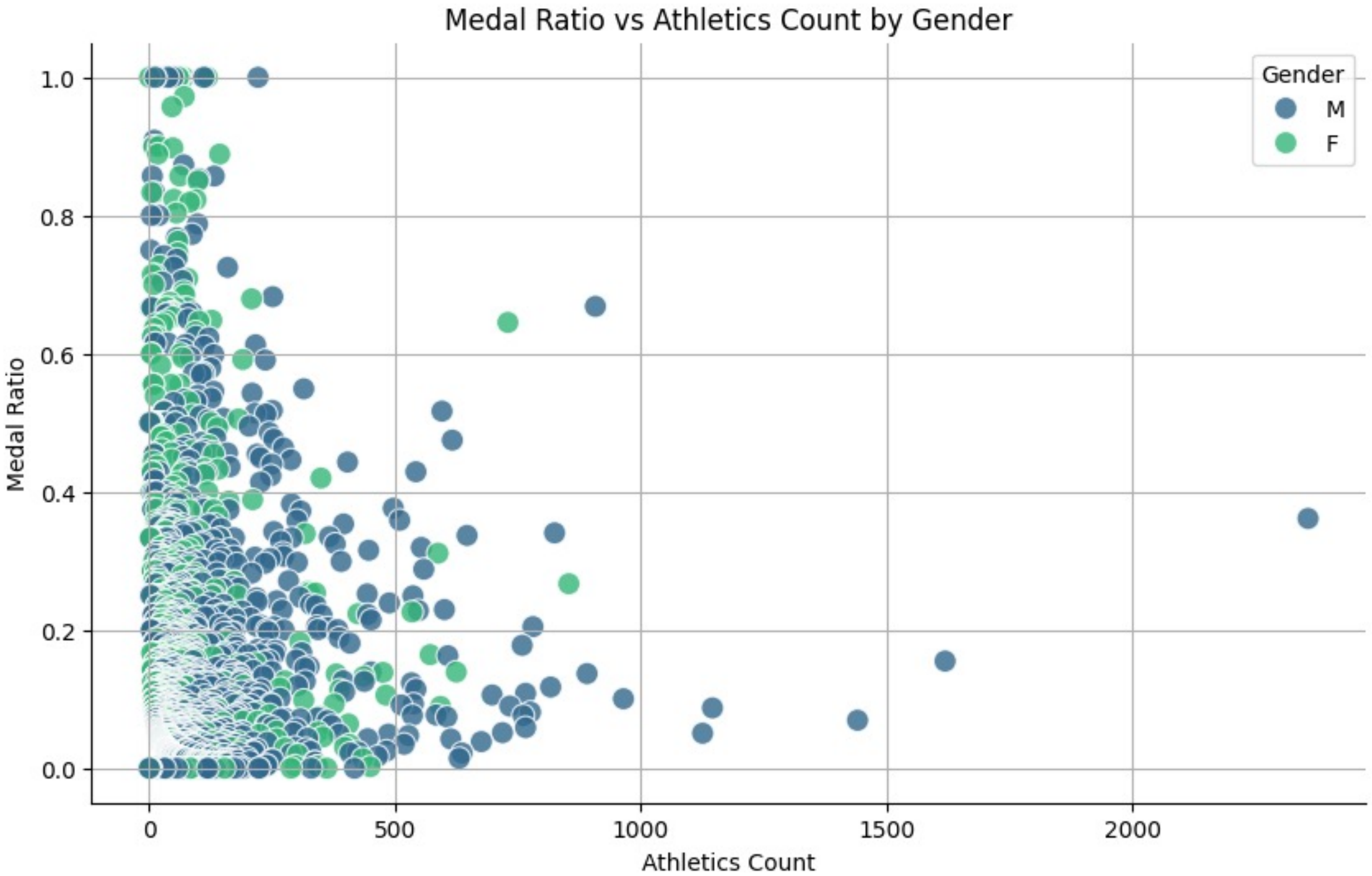
Olympics

Dataset

- The 120 Years of Olympic History Dataset is a historical dataset on the modern Olympic Games, including all games from Athens 1896 to Rio 2016.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 271116 entries, 0 to 271115
Data columns (total 15 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   ID      271116 non-null  int64
 1   Name    271116 non-null  object
 2   Gender  271116 non-null  object
 3   Age     261642 non-null  float64
 4   Height  210945 non-null  float64
 5   Weight  208241 non-null  float64
 6   Team    271116 non-null  object
 7   NOC     271116 non-null  object
 8   Games   271116 non-null  object
 9   Year    271116 non-null  int64
 10  Season  271116 non-null  object
 11  City    271116 non-null  object
 12  Sport   271116 non-null  object
 13  Event   271116 non-null  object
 14  Medal   39783 non-null   object
dtypes: float64(3), int64(2), object(10)
memory usage: 31.0+ MB
```

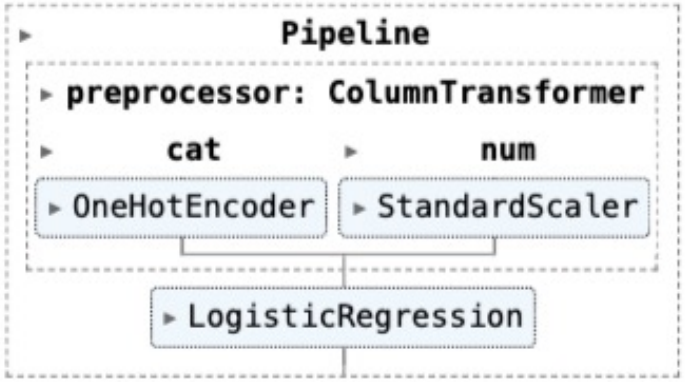

Medals Won Over Time for Different NOCs
NOC
NED
FRA
ITA
USA
ROC
HUN
GER
AUS
SWE
GBR
Number of Medals Won
0
100
200
300
400
Year
1900
1920
1940
1960
1980
2000
2020


Medal Ratio vs Athletics Count by Gender
Gender
M
F
Medal Ratio
1.0
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
Athletics Count

Module1: build up Logistic Regression for perdition (All NOC)

	ID	Name	Gender	Age	Height	Weight	NOC	Year	Sport	Event	Medal	Medal_Won
0	1	A Dijiang	M	24.0	180.00000	80.000000	CHN	1992	Basketball	Basketball Men's Basketball	NaN	0
1	2	A Lamusi	M	23.0	170.00000	60.000000	CHN	2012	Judo	Judo Men's Extra-Lightweight	NaN	0
2	3	Gunnar Nielsen Aaby	M	24.0	175.33897	70.702393	DEN	1920	Football	Football Men's Football	NaN	0
3	4	Edgar Lindenau Aabye	M	34.0	175.33897	70.702393	DEN	1900	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold	1
26	8	Cornelia "Cor" Aalten (-Strannood)	F	18.0	168.00000	70.702393	NED	1932	Athletics	Athletics Women's 100 metres	NaN	0

- Logistic Regression
- features = ['Gender', 'Age', 'Height', 'Weight', 'NOC', 'Sport']
- target = 'Medal_Won'



Accuracy: 0.85

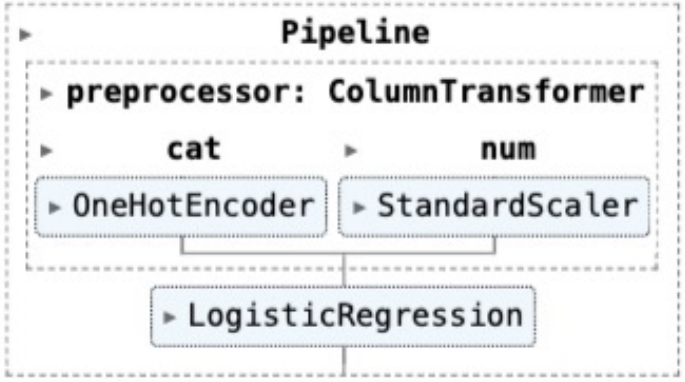
	precision	recall	f1-score	support
0	0.86	0.99	0.92	37703
1	0.58	0.09	0.15	6808
accuracy			0.85	44511
macro avg	0.72	0.54	0.53	44511
weighted avg	0.81	0.85	0.80	44511

```
[[37283   420]
 [ 6224   584]]
```

Module2: build up Logistic Regression for perdition (USA)

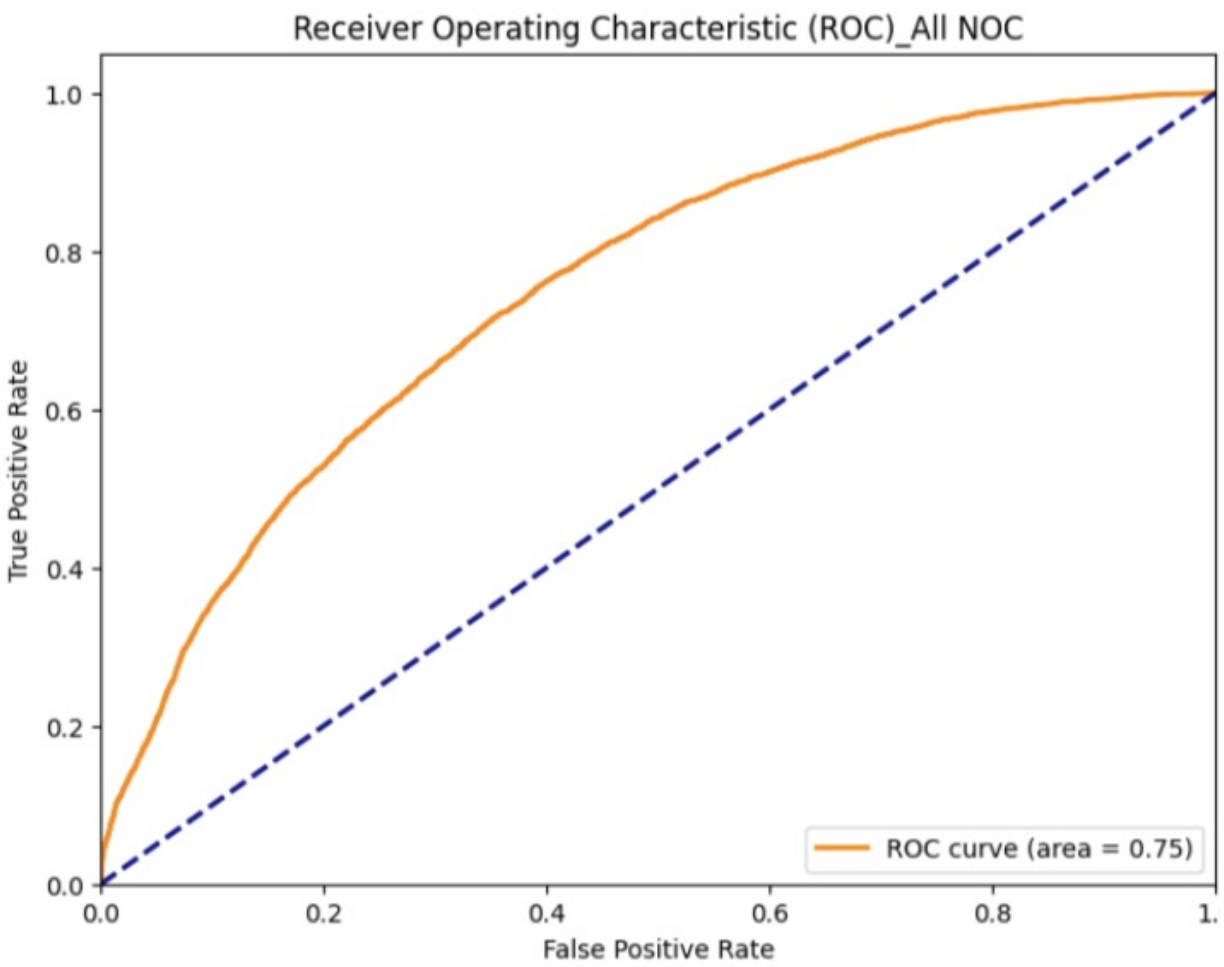
	ID	Name	Gender	Age	Height	Weight	NOC	Year	Sport	Event	Medal	Medal_Won
186	84	Stephen Anthony Abas	M	26.0	165.00000	55.000000	USA	2004	Wrestling	Wrestling Men's Featherweight, Freestyle	Silver	1
273	142	David "Dave" Abbott	M	26.0	183.00000	75.000000	USA	1928	Athletics	Athletics Men's 5,000 metres	NaN	0
282	149	Mara Katherine Abbott	F	30.0	163.00000	52.000000	USA	2016	Cycling	Cycling Women's Road Race, Individual	NaN	0
283	150	Margaret Ives Abbott (–Dunne)	F	23.0	175.33897	70.702393	USA	1900	Golf	Golf Women's Individual	Gold	1
284	151	Mary Perkins Ives Abbott (Perkins–)	F	42.0	175.33897	70.702393	USA	1900	Golf	Golf Women's Individual	NaN	0

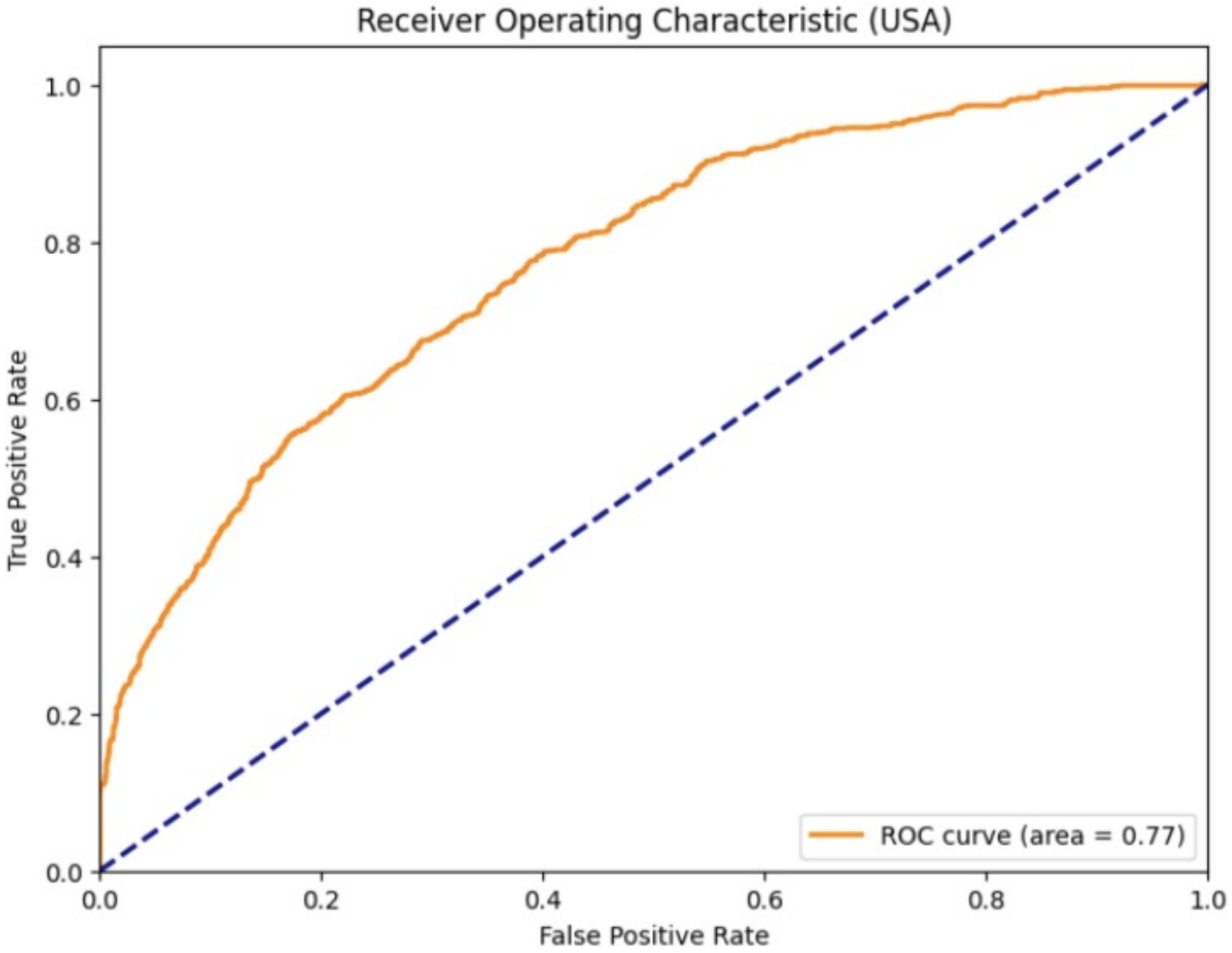
- Logistic Regression
- features = ['Gender', 'Age', 'Height', 'Weight', 'Sport']
- target = 'Medal_Won'




```
Accuracy: 0.74
              precision    recall  f1-score   support

           0       0.74      0.93      0.83      2026
           1       0.71      0.34      0.46       987

    accuracy                           0.74      3013
   macro avg       0.73      0.64      0.65      3013
weighted avg       0.73      0.74      0.71      3013

[[1890  136]
 [ 647  340]]
```

ROC charts for module1 and module2





2021 Tokyo Olympics Team USA Data Analytics

Olympics

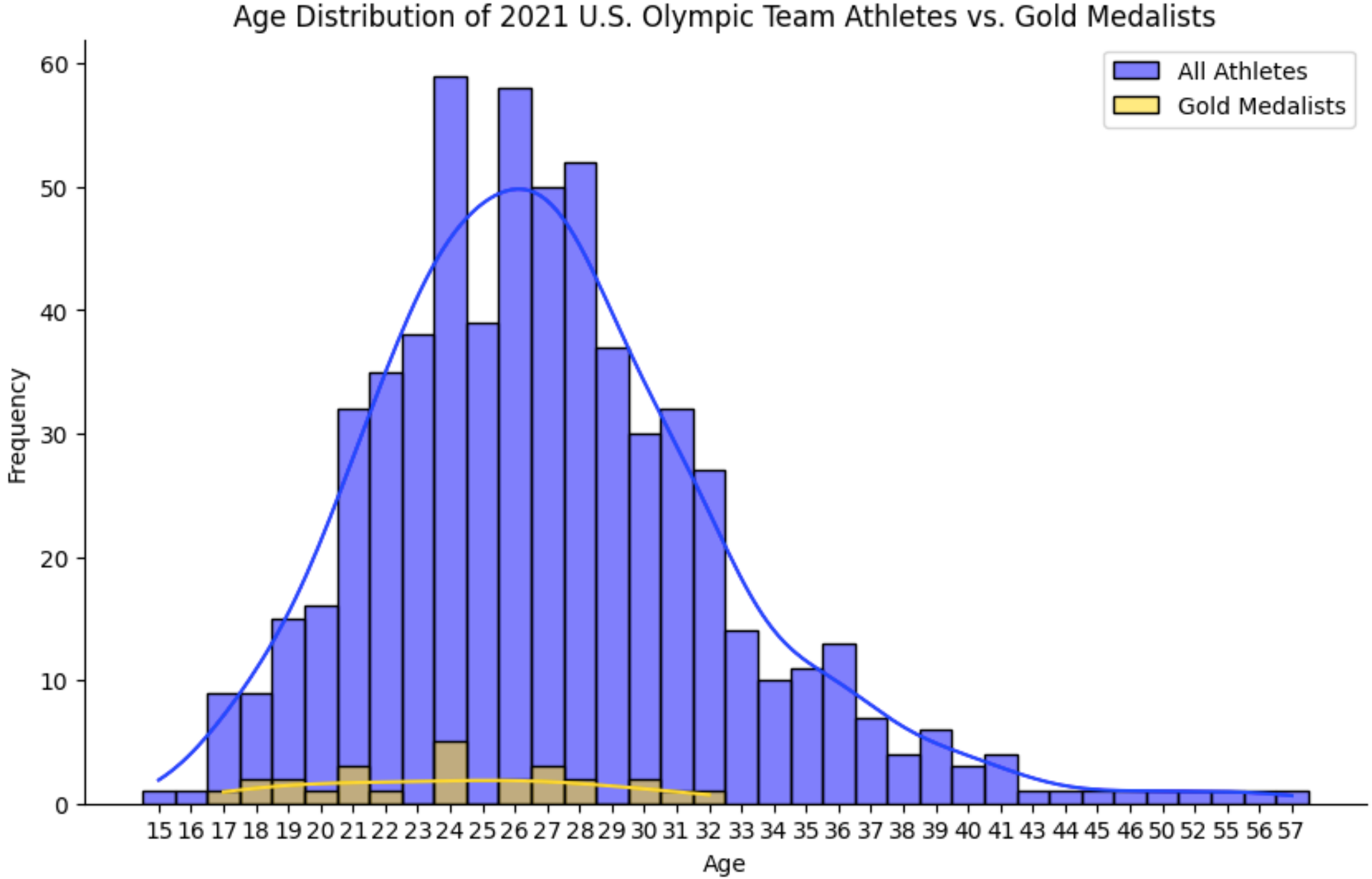
Dataset

Dataset 1: Team USA participated

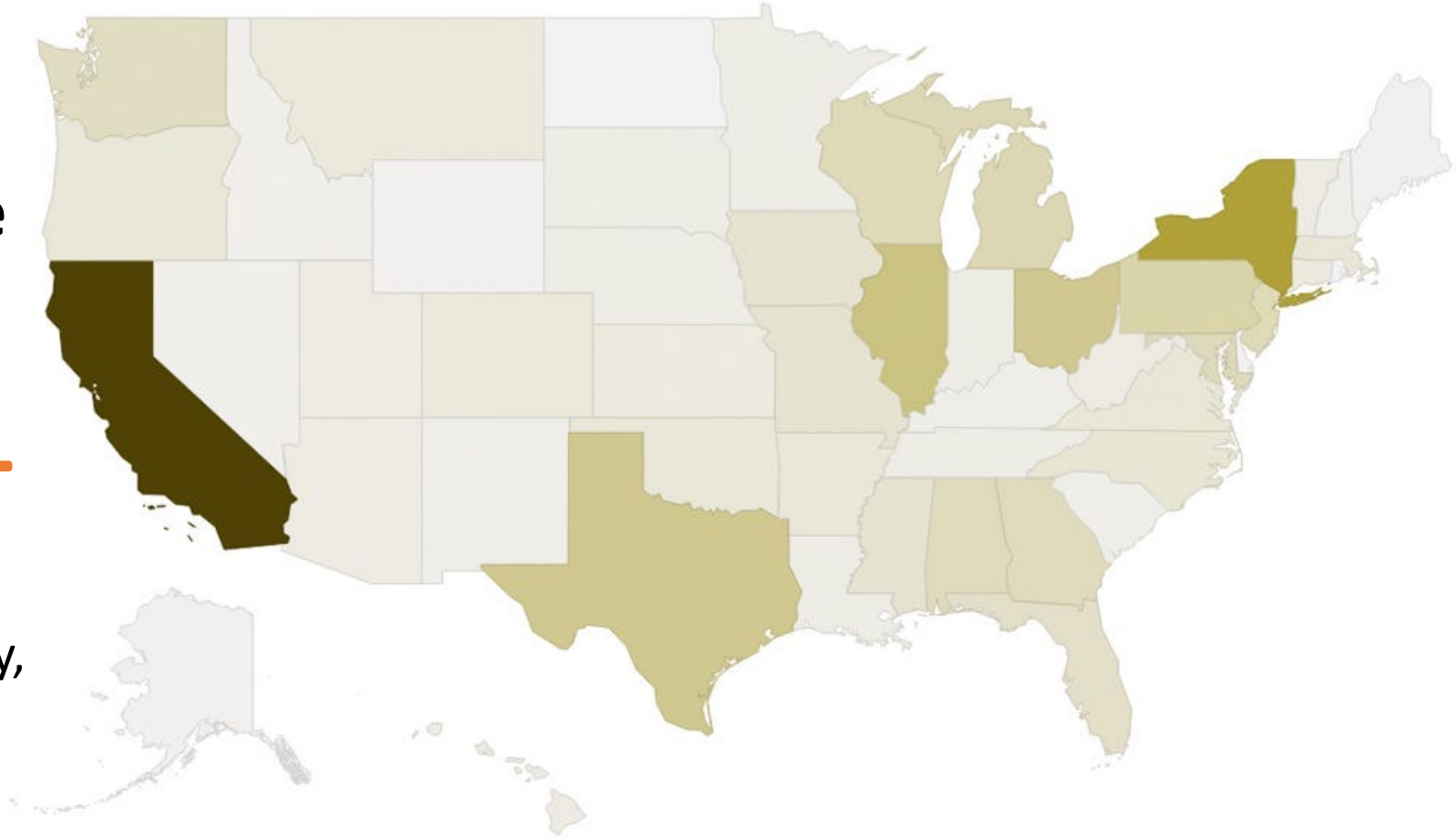
	Athlete	Hometown	Birthplace	Age
0	Mackenzie Brown	Flint, Texas	Flint, Texas	26
1	Brady Ellison	Globe, Ariz.		32
2	Casey Kaufhold	Lancaster, Pa.		17
3	Jennifer Mucino-Fernandez	Mexico City, Mexico	Boston	18
4	Jack Williams	Irvine, Calif.	Irvine, Calif.	21
5	Jacob Wukie	Oak Harbor, Ohio	Massillon, Ohio	35
6	Phillip Chew	Orange, Calif.	Anaheim, Calif.	27
7	Ryan Chew			24
8	Timothy Lam			23
9	Beiwen Zhang	Las Vegas		31

Dataset 2: Team USA who won medals

	Medal	Name	Sport	Event
0	Gold	Lee Kiefer	Fencing	Women's foil
1	Gold	Will Shaner	Shooting	Men's 10 m air rifle
2	Gold	Chase Kalisz	Swimming	Men's 400 m individual medley
3	Gold	Anastasija Zolotic	Taekwondo	Women's −57 kg
4	Gold	Zach AppleBowe BeckerBrooks Curry[a]Caeleb Dre...	Swimming	Men's 4 × 100 m freestyle relay
5	Gold	Vincent Hancock	Shooting	Men's skeet
6	Gold	Amber English	Shooting	Women's skeet
7	Gold	Carissa Moore	Surfing	Women's shortboard
8	Gold	Lydia Jacoby	Swimming	Women's 100 m breaststroke
9	Gold	United States women's national 3x3 teamStefani...	Basketball	Women's 3x3 tournament


Age Distribution of 2021 U.S. Olympic Team Athletes vs. Gold Medalists
All Athletes
Gold Medalists
Frequency
0
10
20
30
40
50
60
15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 43 44 45 46 50 52 55 56 57
Age

The States With The Most Olympic Gold Medals 1924 - 2014



- California is the most athletic state in the country, Wisconsin punches above its weight, and Florida is slacking.
- Business Insider only counted individual Olympic gold medals

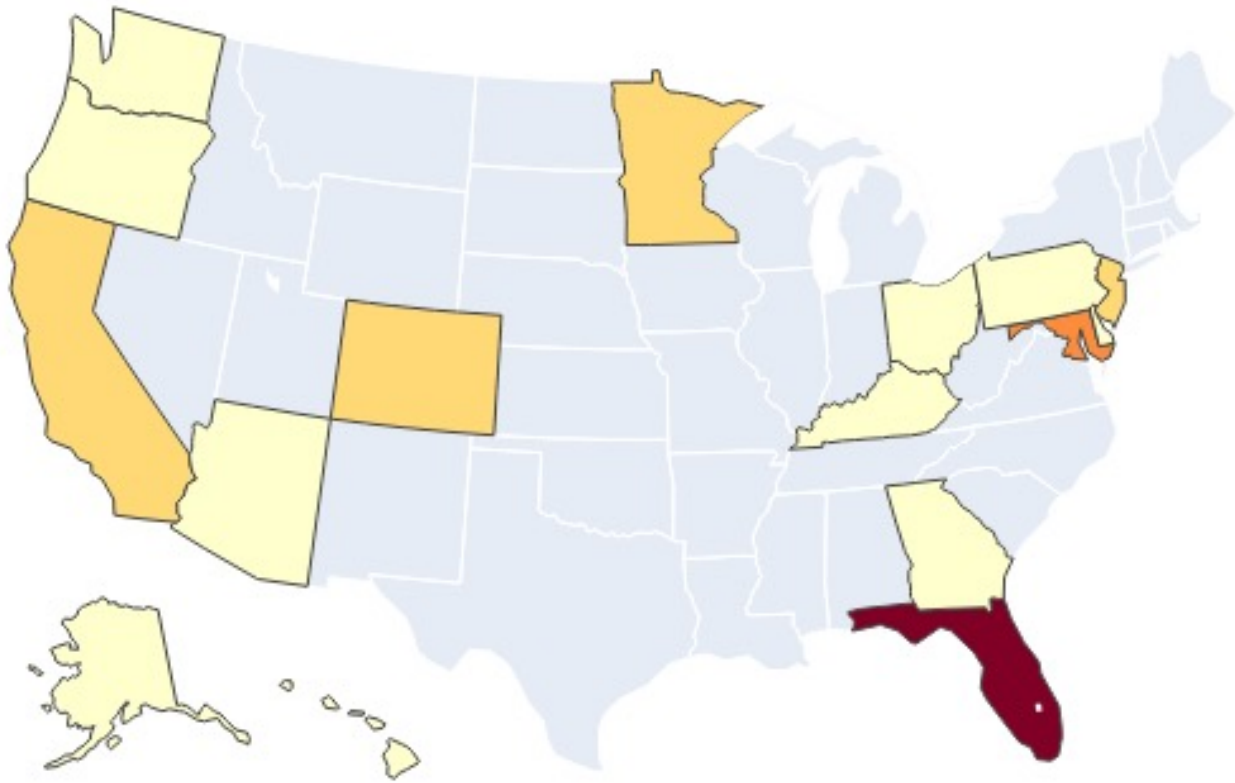
1	California	122
2	New York	61
3	Illinois	37
4	Ohio	32
5	Texas	32
6	Pennsylvania	23
7	Michigan	20
8	Wisconsin	19
9	New Jersey	19
10	Maryland	18
11	Georgia	17
12	Alabama	17
13	Washington	16
14	Florida	13

15	Iowa	11
16	Mississippi	11
17	Missouri	11
18	North Carolina	10
19	Massachusetts	9
20	Virginia	9
21	Oregon	8
22	Arkansas	8
23	Oklahoma	8
24	Montana	7
25	Colorado	7
26	Connecticut	6
27	Washington, DC	6
28	Kansas	6

29	Vermont	5
30	Utah	5
31	Arizona	5
32	West Virginia	5
33	Hawaii	5
34	Minnesota	4
35	Tennessee	4
36	Louisiana	4
37	Nebraska	4
38	South Dakota	4
39	Indiana	4
40	New Hampshire	3
41	New Mexico	3
42	Kentucky	3

43	South Carolina	3
44	Nevada	2
45	Idaho	2
46	Wyoming	1
47	Maine	1
48	Alaska	1
49	Delaware	0
50	North Dakota	0
51	Rhode Island	0

Gold Medal Winners Hometowns in USA

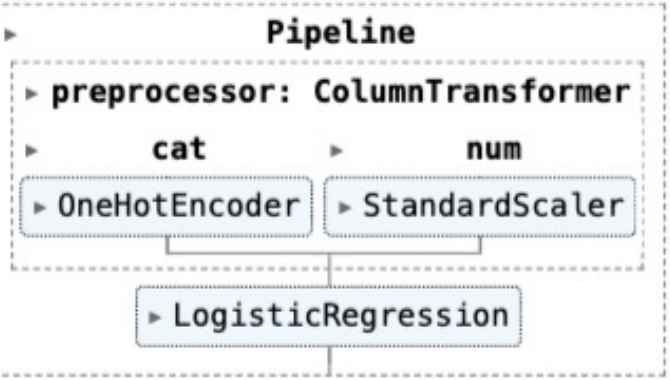


Number of Gold Medal Winners Hometown

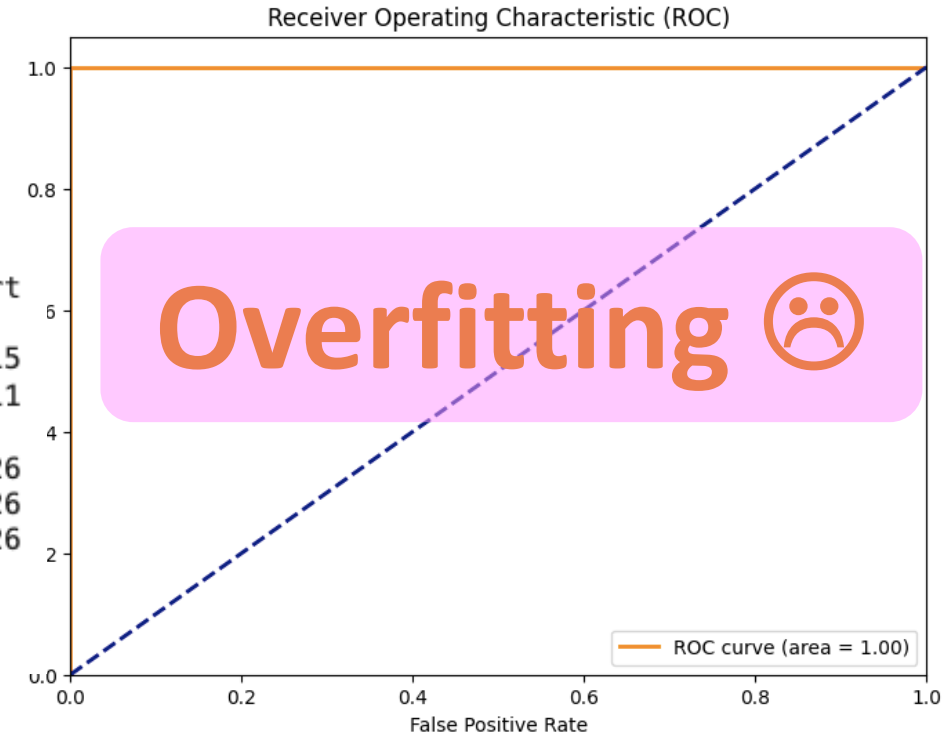
5
4.5
4
3.5
3
2.5
2
1.5
1

Module3: build up Logistic Regression for perdition (USA 2021)

- `Logistic Regression`
- `features = ['Age', 'Hometown', 'Sport']`
- `target = 'Medal_Won'`




```
Accuracy: 1.00
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       115
           1       1.00      1.00      1.00        11

    accuracy                           1.00       126
   macro avg       1.00      1.00      1.00       126
weighted avg       1.00      1.00      1.00       126

[[115   0]
 [  0  11]]
```

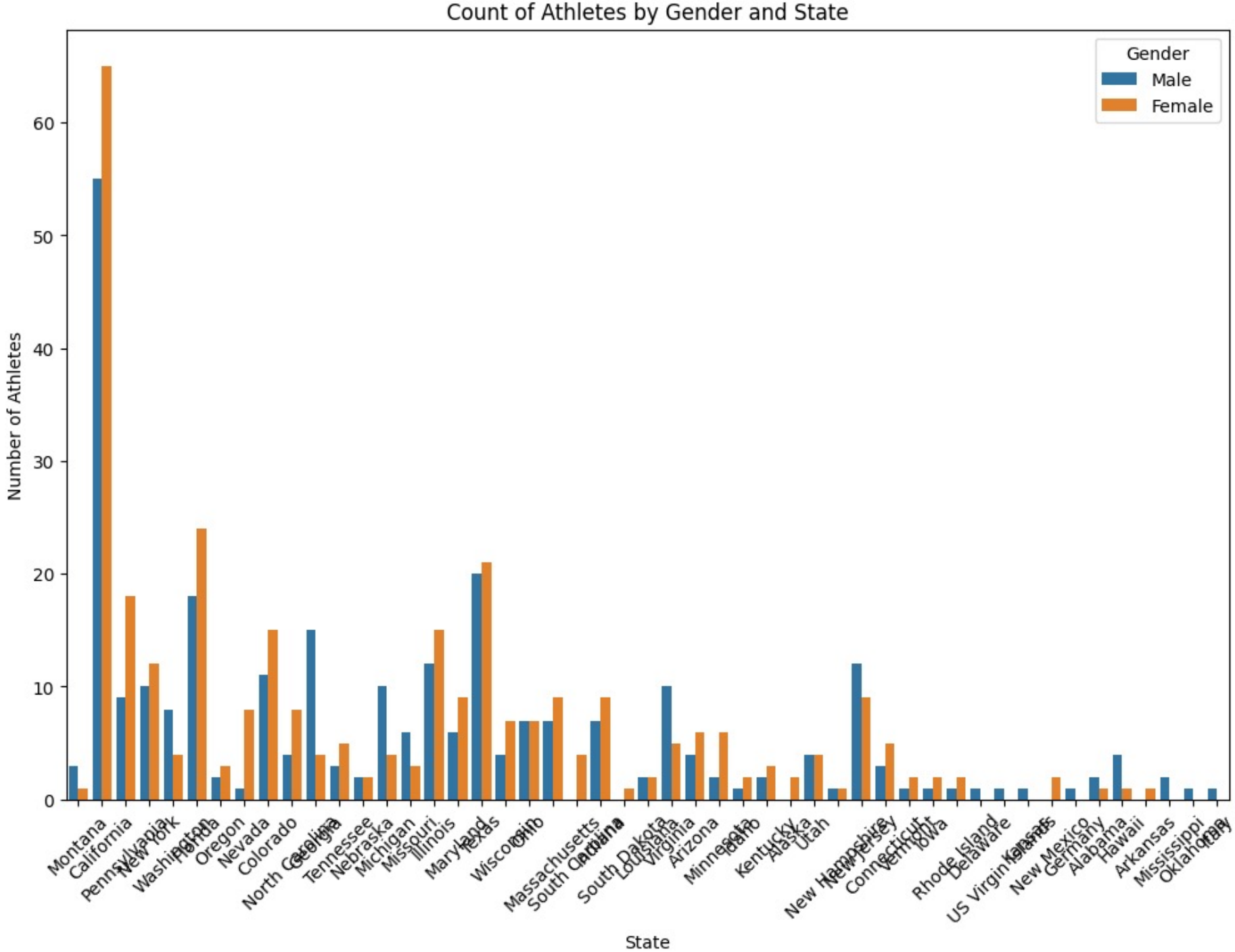


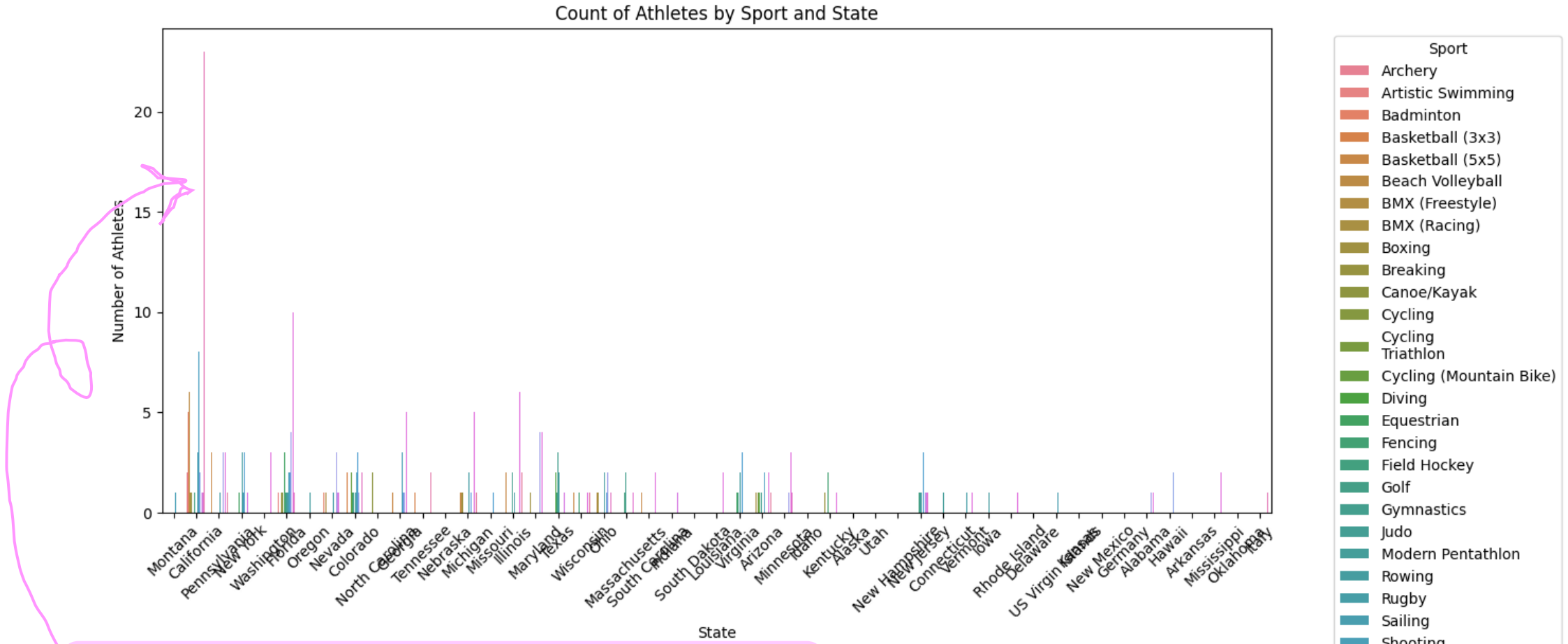
2024 Paris Olympics Team USA Data Analytics

Olympics

Dataset

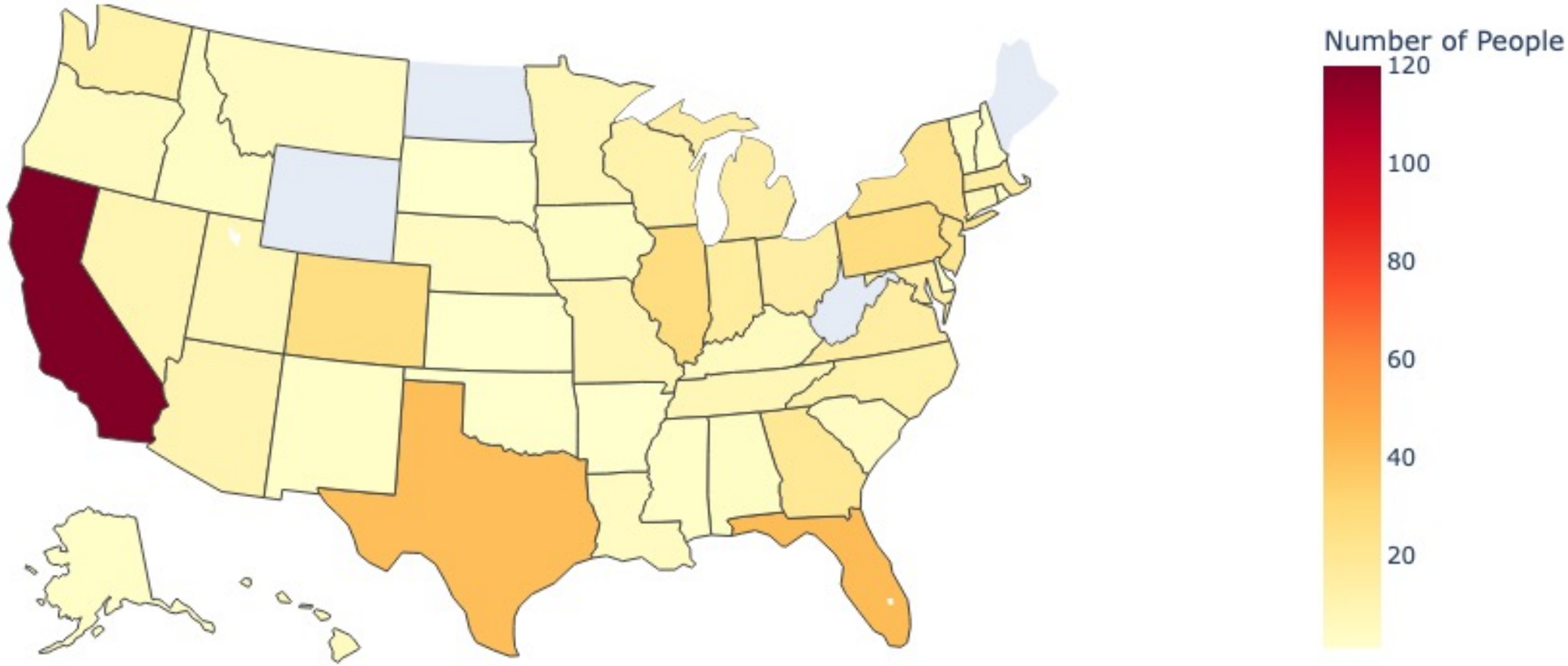
	First Name	Last Name	Sport	Hometown City	Hometown State	Gender	Event
0	BRADY	ELLISON	Archery	Billings	Montana	Male	Men's Individual
1	CATALINA	GNORIEGA	Archery	San Diego	California	Female	Women's Individual, Women's Team
2	CASEY	KAUFHOLD	Archery	Lancaster	Pennsylvania	Female	Women's Individual, Women's Team
3	JENNIFER	MUCINO-FERNANDEZ	Archery	Chula Vista	California	Female	Women's Individual, Women's Team
4	ANITA	ALVAREZ	Artistic Swimming	Buffalo	New York	Female	Team


Count of Athletes by Gender and State
Gender
Male
Female
Number of Athletes
0
10
20
30
40
50
60
State
Montana
California
Oregon
Nevada
Colorado
Tennessee
Nebraska
Michigan
Missouri
Illinois
Texas
Massachusetts
South Dakota
Virginia
Arizona
Minnesota
Kentucky
Alaska
Utah
New Jersey
Vermont
Iowa
Rhode Island
Delaware
New Mexico
Germany
Alabama
Hawaii
Arkansas
Mississippi



California Water Polo 23

Number of People by Hometown State



Conclusion

- The Module 2 (using 120 years of historic data `'Gender', 'Age', 'Height', 'Weight', 'Sport'`) is the best, with 74% accuracy and ROC area 0.77
- Using only 2021 Olympics data to build up module is hard to predict an efficiently result
- If I have more time to collect all of the features above for every athletes participating 2024 Olympics, I can predict who could win a medal this summer.

Take Away

- Olympic data are a big tricky, some of the country changed their name from time to time (Russia > ROC), and lots of group games. I also learnt lots of different sports.
- I learnt that we can getting data from HTML to NumPy by using several packages. I thought apis was the only way before.
- I only used one machine learning technique: Logistic Regression. I would like to use those dataset that I collected to train and tun more module after this competition.

Thank you