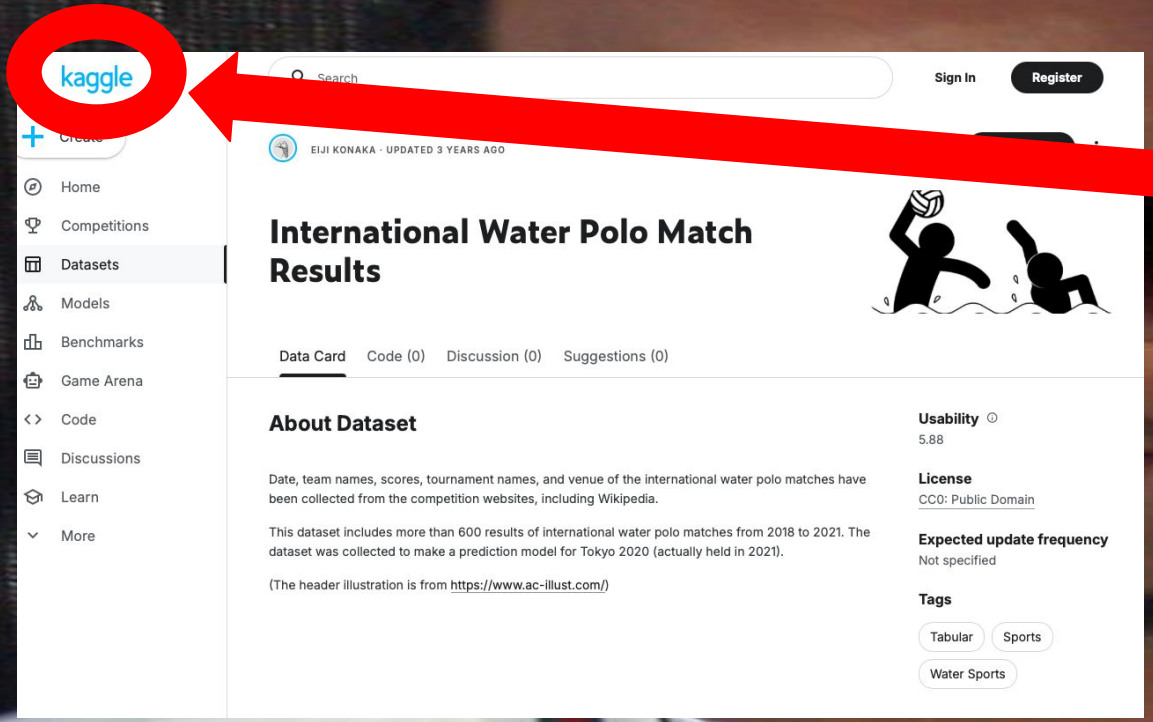
Waterpolo data

By Nikita and Gio

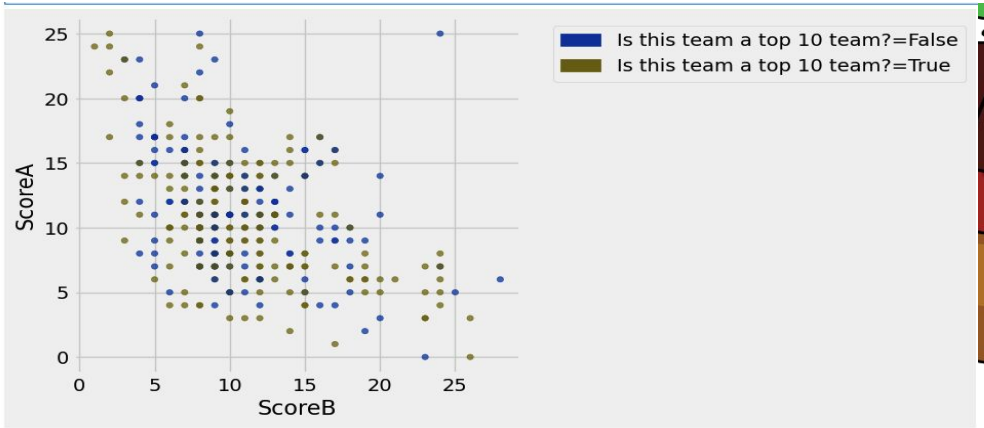
Overview



- The dataset that we used for this project was called International Water Polo Match Results
- It has team stats from each team from 2018 to 2020
Dataset from Kaggle
- Data organized by a user named EIJI KONAKA.
- It includes the tournament names.

Variables

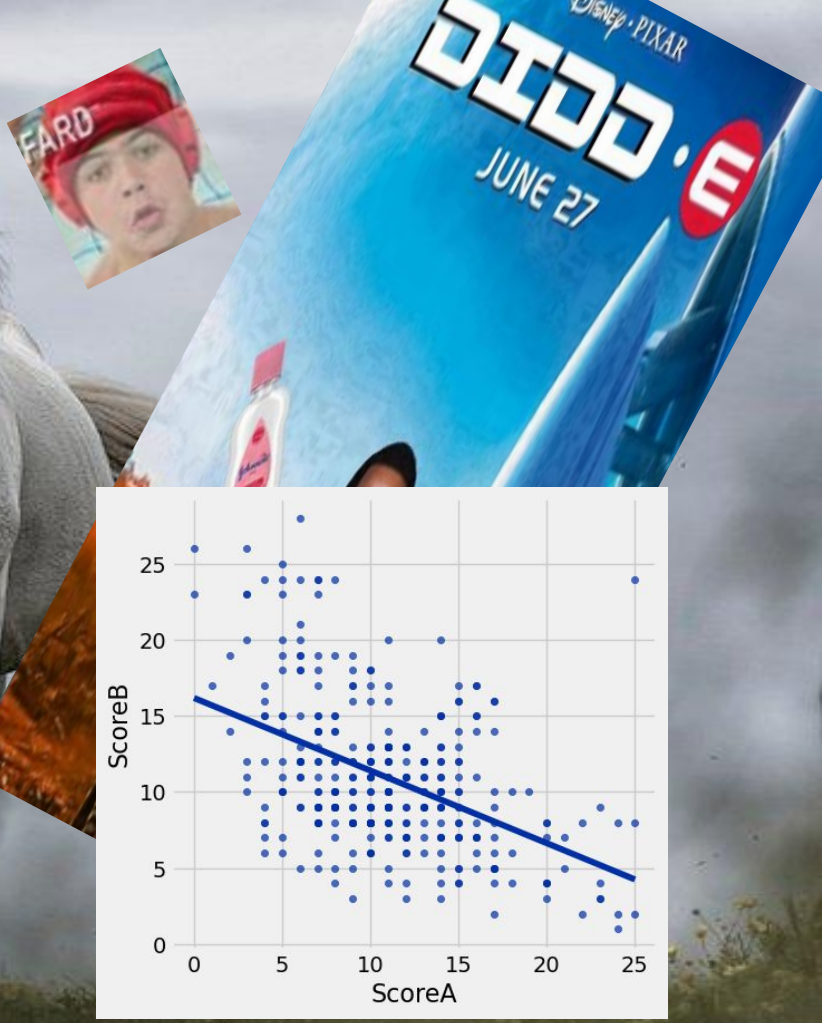
- We are looking to see whether or not the teams are top 10



```
euhgyvkgiuyutvfkylbhijn = make_array('Spain','Crotia','Hungary','Greece','Serbia','Italy','United States','Montenegro','Brazil','Australia')
def WithinTheTop10YesExclamationPoint(aihjsojrgsahrd,bauihgvfoubaghir):
    cguhiglyukcgvylihglyuftk = aihjsojrgsahrd in euhgyvkgiuyutvfkylbhijn
    dubiyuvktybiuyvkucyfvugliyuvtcykuv = bauihgvfoubaghir in euhgyvkgiuyutvfkylbhijn
    return cguhiglyukcgvylihglyuftk or dubiyuvktybiuyvkucyfvugliyuvtcykuv
WithinTheTop10YesExclamationPoint('Spain','Serbia')
my_data_raw = my_data_raw.with_column(
    'Is this team a top 10 team?',
    my_data_raw.apply(WithinTheTop10YesExclamationPoint, 'TeamA','TeamB')
)
my_data_raw.show()
```

Variables Pt.2

- Potential bias = Refs heavily influencing game outcomes
- 95% Correlation Confidence Interval: [-0.5593931042730582, -0.3468940560449595] , Reject the null: True
- We used 325 games as our data
- This shows a correlation where the line goes down
- ScoreA (x) is the score of team A in the dataset at the end of the math.
- ScoreB (y) is the score of the B team (Team against team A) at the end of the math.



Features + Classifier

Date	TeamA	TeamB	ScoreA	ScoreB	Sex	TournamentName	year	Venue	Team A won	Is this team a top 10 team?	Predictions	Was correct
2021/2/17	Croatia	Russia	13	14	M	OlympicQ	2021	Netherlands	False	False	True	False
2020/1/24	Turkey	Romania	3	20	M	EuropeanChampionship	2020	Hungary	False	False	True	False
2019/2/19	Serbia	Romania	13	6	M	WaterPoloWorldLeague	2019	Serbia	True	True	False	False
2021/2/19	Georgia	Croatia	6	15	M	OlympicQ	2021	Netherlands	False	False	True	False
2019/7/17	Hungary	Spain	13	11	M	WorldChampionship	2019	South Korea	True	True	False	False
2019/8/6	United States	Puerto Rico	24	1	M	PanAmericanGames	2019	Peru	True	True	False	False
2019/7/23	Kazakhstan	Brazil	8	9	M	WorldChampionship	2019	South Korea	False	True	False	False
2019/11/25	Spain	Ukraine	20	8	M	WaterPoloWorldLeague	2020	Spain	True	True	False	False
2019/3/31	Japan	Australia	8	10	M	WaterPoloWorldLeague	2019	Australia	False	True	False	False
2020/1/22	Germany	Romania	15	10	M	EuropeanChampionship	2020	Hungary	True	False	True	False
2018/8/29	Singapore	South Korea	7	10	M	AsianGames	2018	Indonesia	False	False	True	False
2020/1/24	Montenegro	Hungary	8	10	M	EuropeanChampionship	2020	Hungary	False	True	False	False
2019/4/7	Italy	Spain	7	9	M	WaterPoloWorldLeague	2019	Croatia	False	True	False	False
2021/2/18	France	Romania	16	7	M	OlympicQ	2021	Netherlands	True	False	True	False
2019/3/29	South Africa	Kazakhstan	8	14	M	WaterPoloWorldLeague	2019	Australia	False	False	True	False

... (66 rows omitted)

The classifier we used was Was it in the top ten teams? We used this to determine the points they would score in a game using an x value of 20. Our accuracy was 58%

After we increased the number of points that we found the distance to from 3 to 9 our accuracy increased slightly to 60%. Due to our tiny limited dataset we couldn't exactly make a whole new classifier

(The process of classification)

KNN (K-nearest

re
and
m

Conclusion

- Wanted to see if games, score and rating are correlated
- Dataset may not be the best for this as there is very little data
- Learned a lot of high-rated teams have fewer games and more games relative to others
- Shows subtle pattern in dataset
- Not great for prediction and classification
- 58% accuracy
- Not practical – Probably easier to just look up the score of a game

Credits