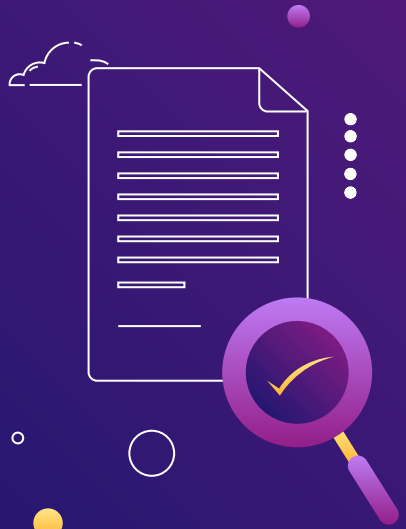
ICPC : EXPLORATORY DATA ANALYSIS & PREDICTION



HELLO!

amr.mail18@gmail.com

TABLE OF CONTENT

01 PROJECT INTRODUCTION

Data set definition and understanding

02 PROBLEMS AND QUESTIONS

Questions about the data

03 DATA UNDERSTANDING

Exploratory data analysis and visuals

04 MODEL AND EVALUATION

Model used on data and the evaluation result

01

PROJECT INTRODUCTION

Data set here

PROJECT PROCESS

PROBLEM DEFINITION
Define problem and questions

01

02

LOAD DATA
Import data set and load it

03

DATA UNDERSTANDING
Look at data features and info

04

EDA
Perform data analysis on the data set

05

PREPROCESSING
Fix data problems

06

MODELING AND EVALUATION
Perform a **ML** model and evaluate it

TOOLS AND LIBRARIES

PYTHON

PANDAS

DATASET INFO

- The ICPC World Finals Ranking Dataset provides information about the performance of teams from different universities around the world which participated in the International Collegiate Programming Contest (ICPC) World Finals from 1999 onwards.
- The ICPC is one of the most prestigious programming competitions in the world, where teams of three students compete to solve algorithmic problems in a time-restricted environment. It is organized by the Association for Computing Machinery (ACM), and the competition has several rounds before the finals, including regional competitions and online contests.

02

PROBLEMS & QUESTIONS

UNDERSTANDING DATASET

PROBLEMS

This dataset can be used to perform statistical analyses, trend analyses and machine learning models.

QUESTIONS

1. which country hosts the competition the most ?
2. which university is the most participated in the competition ?
3. what is the average number of teams every year ?
4. which university won the gold prize the most ?
5. which university won the silver prize the most ?
6. which university won the bronze prize the most ?
7. what is the average score scored by teams ?
8. what is the average penalty ?
9. is there a relation between the total number of problems and the score achieved ?
10. is there a relation between the time of competition and the champion ?
11. is there a relation to predict when the competition will be hold

03

DATA UNDERSTANDING

DATA COLUMNS AND DESCRIPTION

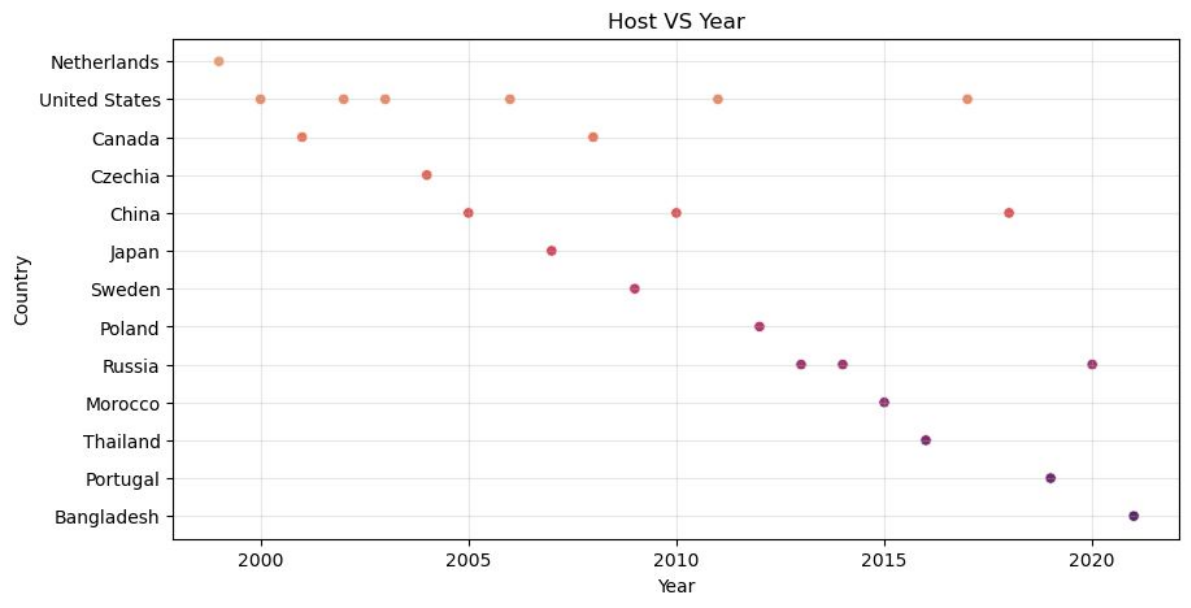
```
Index(['Year', 'Date', 'Host', 'City', 'Venue', 'Rank', 'University',
       'Country', 'Team', 'Contestant 1', 'Contestant 2', 'Contestant 3',
       'Gold', 'Silver', 'Bronze', 'Honorable', 'Score', 'Total',
       'Score Percentage', 'Penalty', 'Prize'],
      dtype='object')
```

	Year	Score	Total	Score Percentage	Penalty
count	2308.000	2275.000	2308.000	2275.000	2190.000
mean	2011.682	3.876	11.133	0.349	655.083
std	6.320	2.226	1.608	0.195	391.896
min	1999.000	0.000	8.000	0.000	0.000
25%	2007.000	2.000	10.000	0.200	365.250
50%	2012.000	4.000	11.000	0.330	638.000
75%	2017.000	5.000	12.000	0.500	934.500
max	2021.000	13.000	15.000	1.000	1912.000

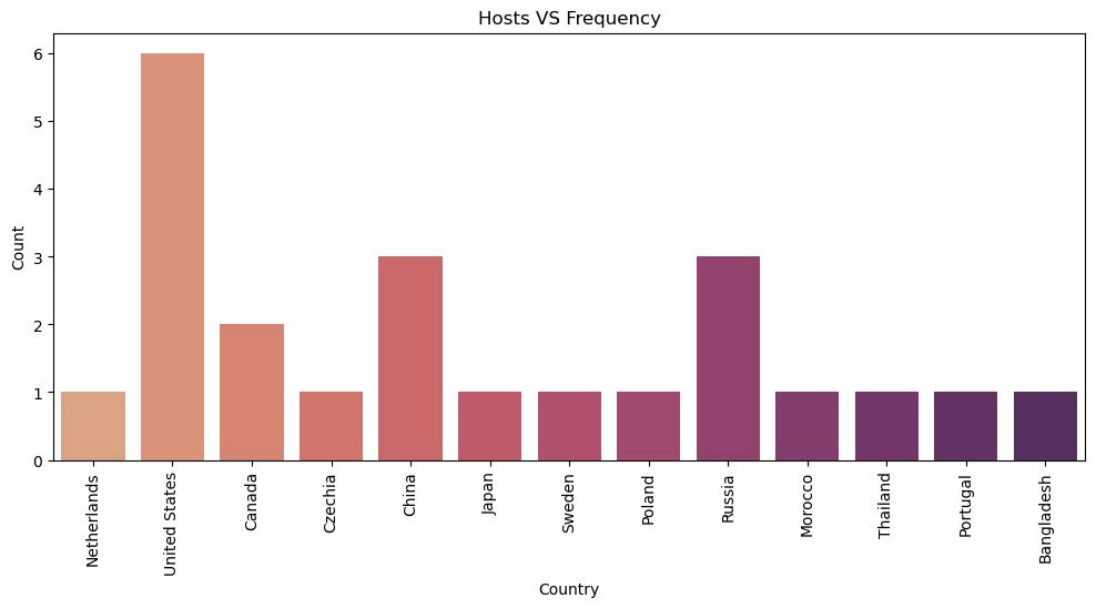
A PICTURE IS WORTH A THOUSAND WORDS

VISUALS

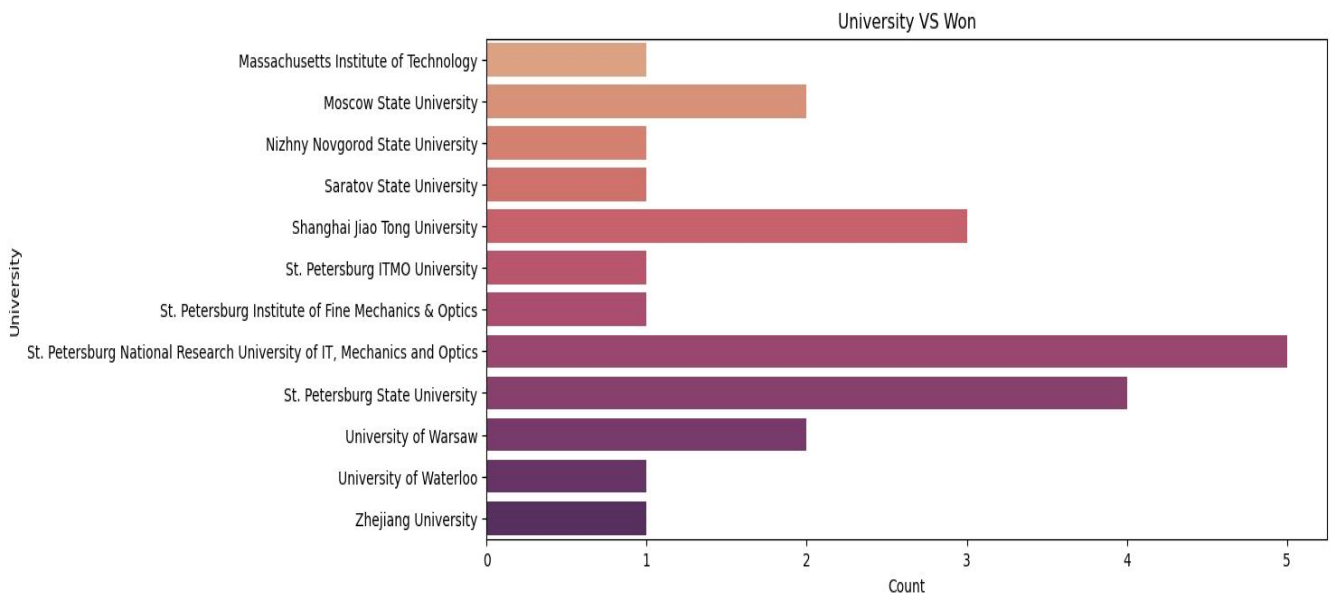
HOST VS YEAR


Host VS Year
Netherlands
United States
Canada
Czechia
China
Japan
Sweden
Poland
Russia
Morocco
Thailand
Portugal
Bangladesh
Country
2000
2005
2010
2015
2020
Year

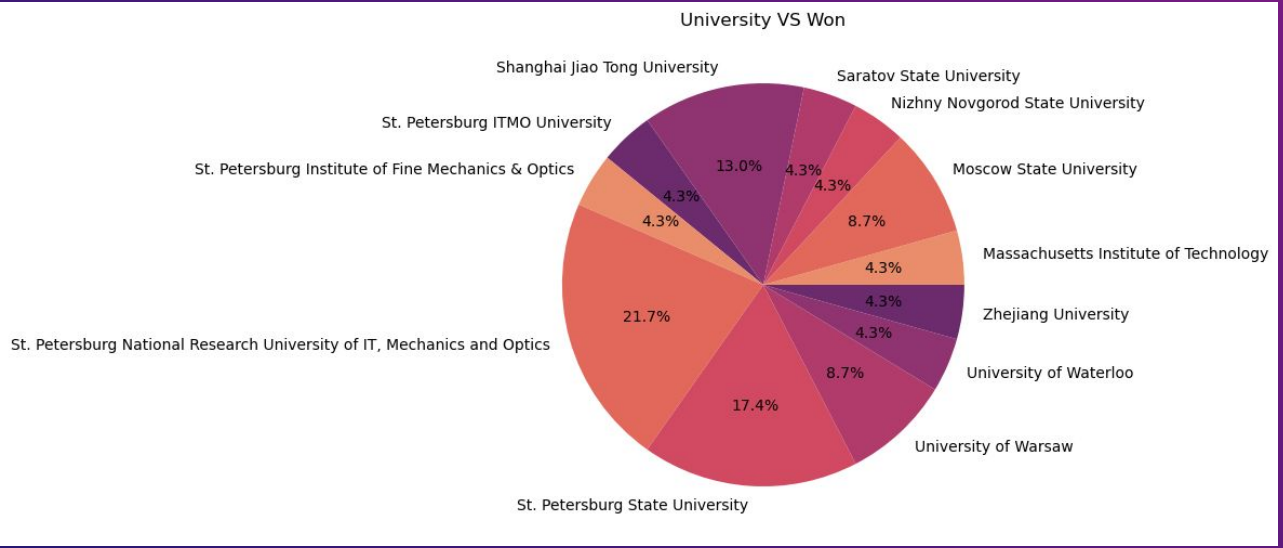
HOSTS VS FREQUENCY



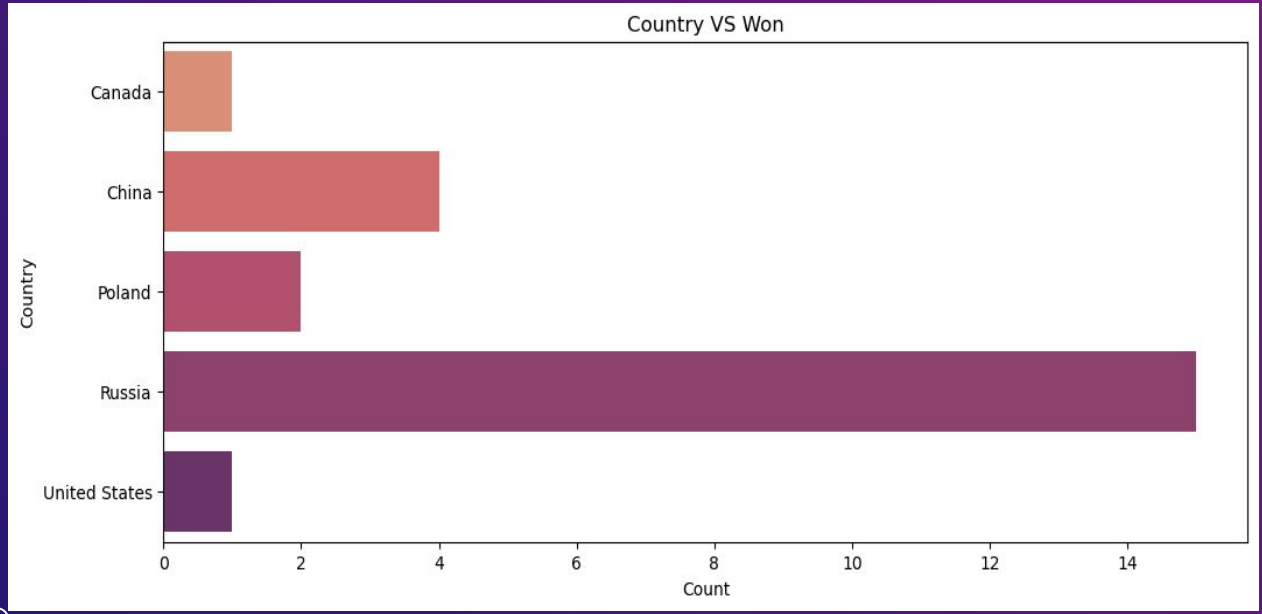
UNIVERSITY VS WON



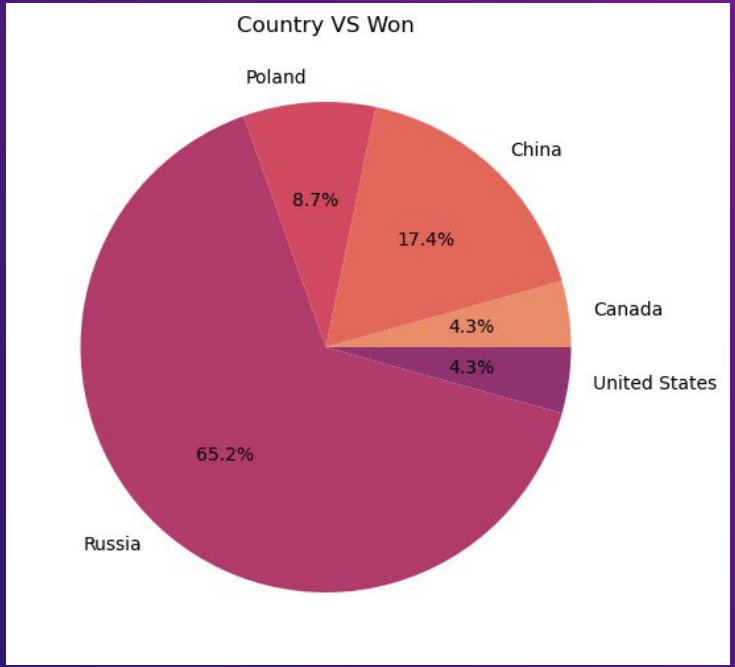
UNIVERSITY VS WON



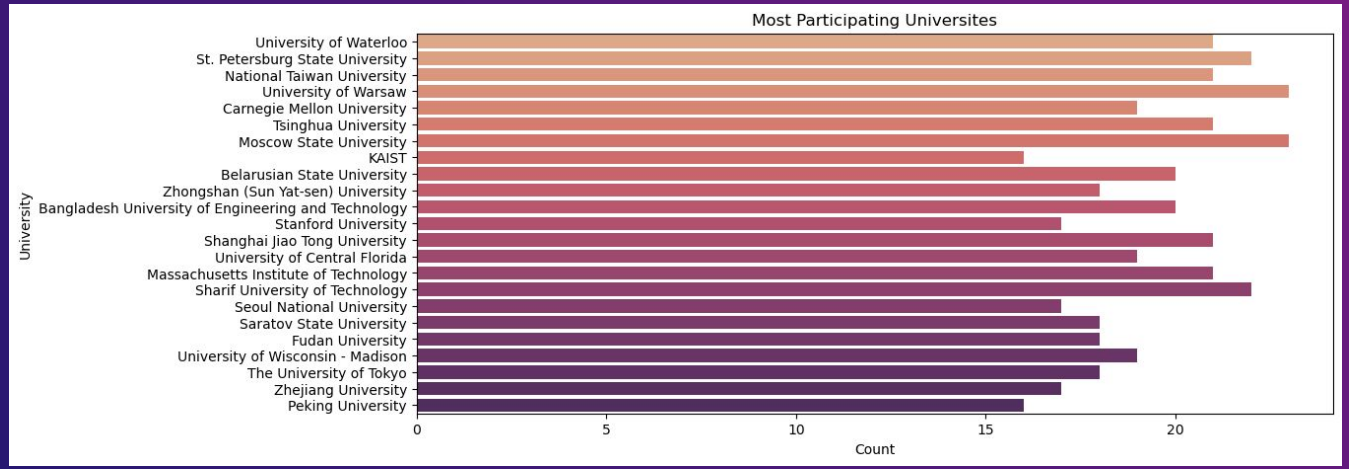
COUNTRY VS WON



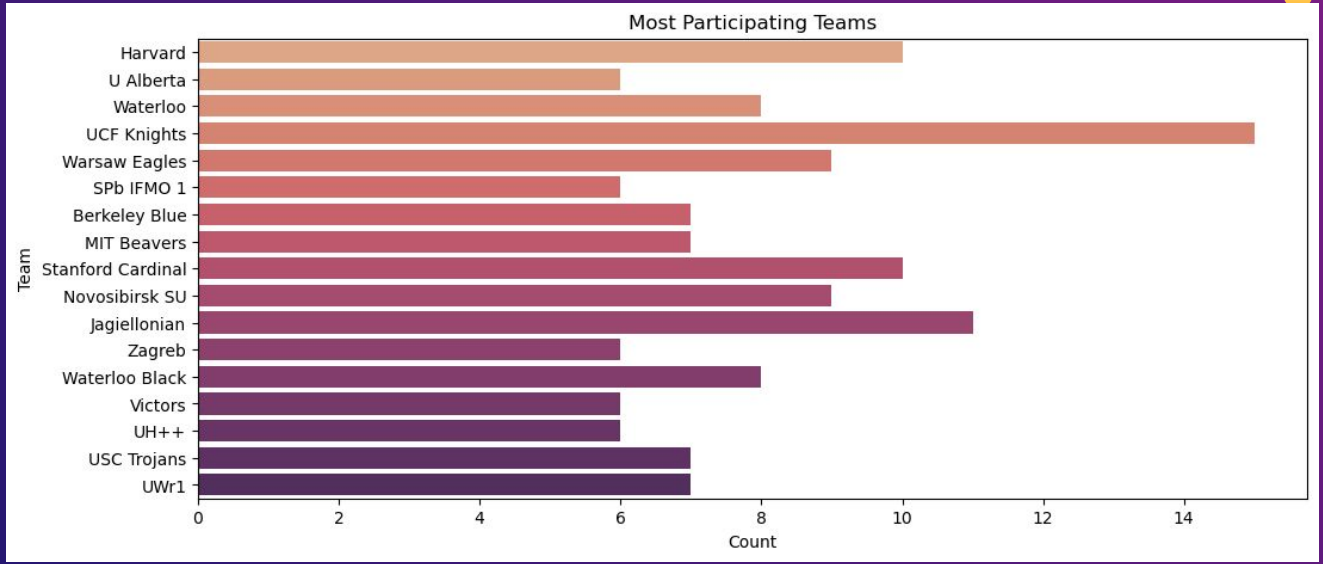
COUNTRY VS WON



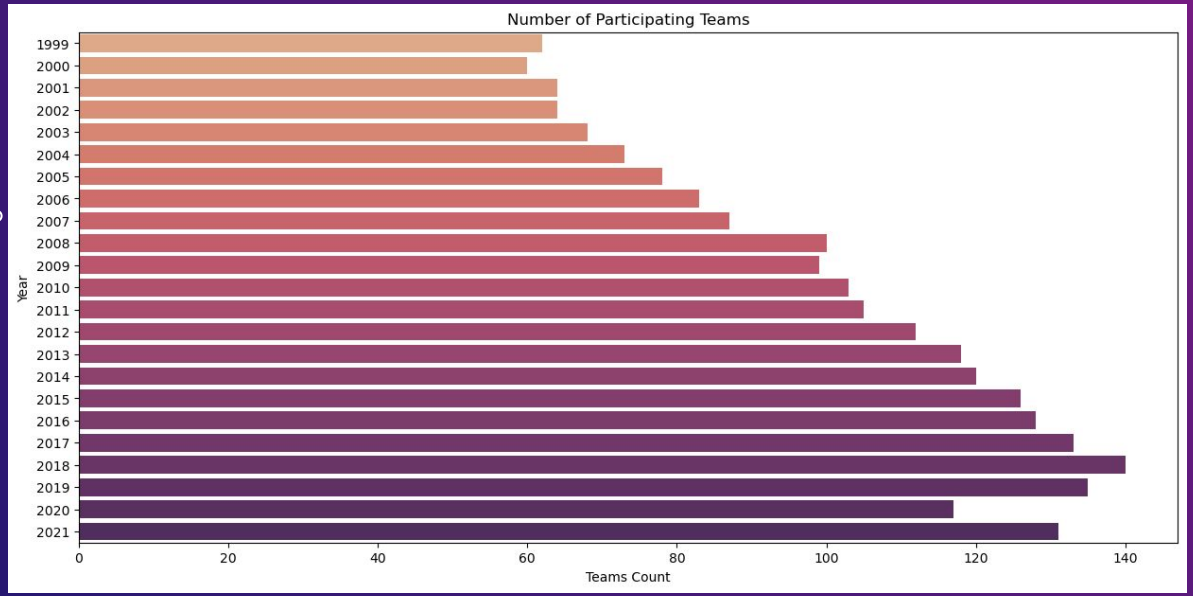
MOST PARTICIPATING UNIVERSITIES



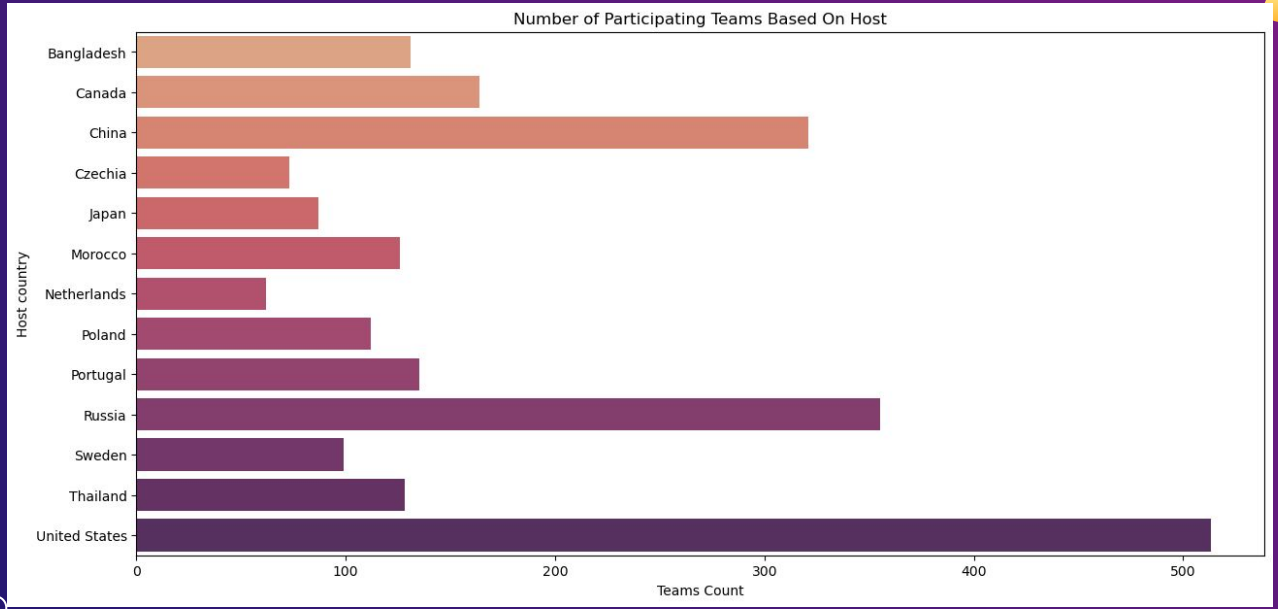
MOST PARTICIPATING TEAMS



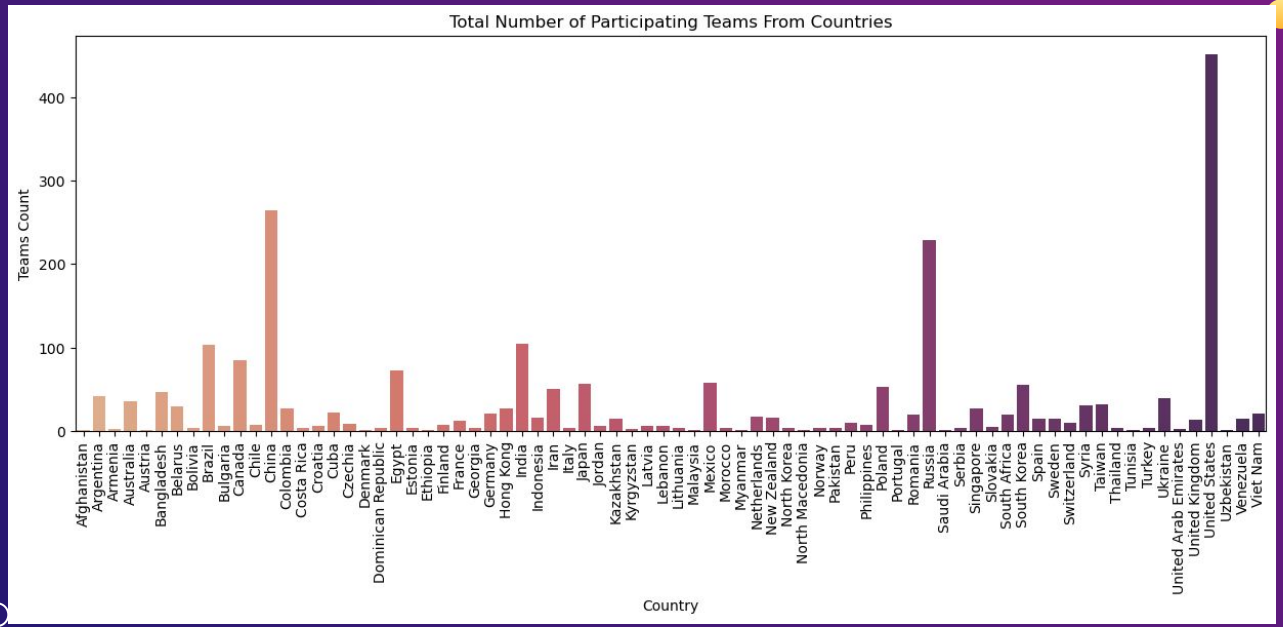
NUMBER OF PARTICIPATING TEAMS



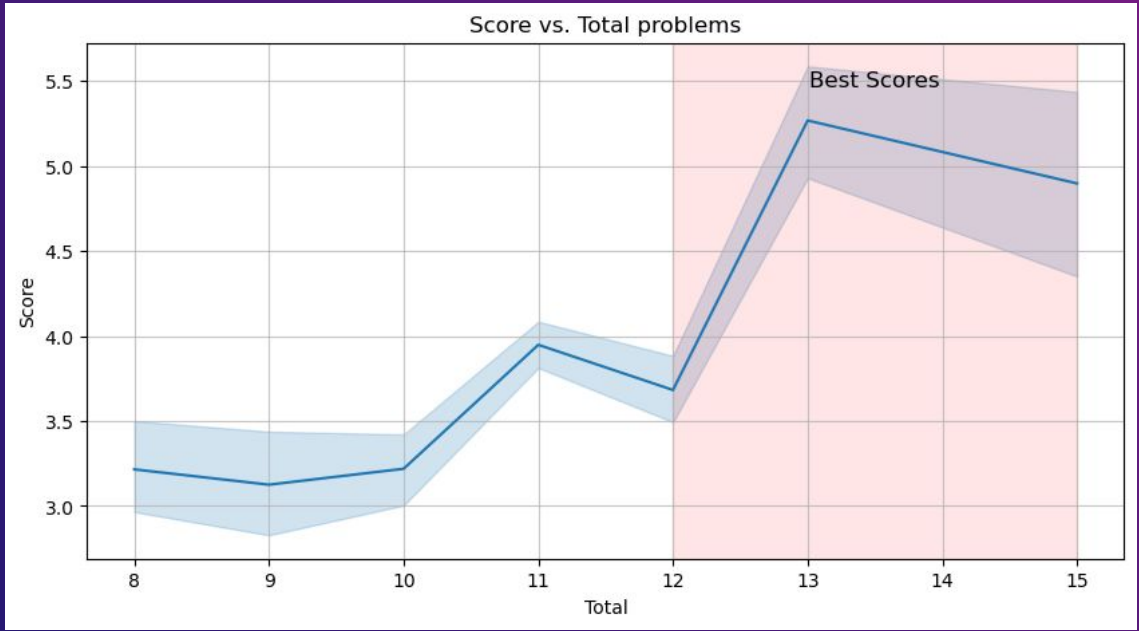
NUMBER OF PARTICIPATING TEAMS



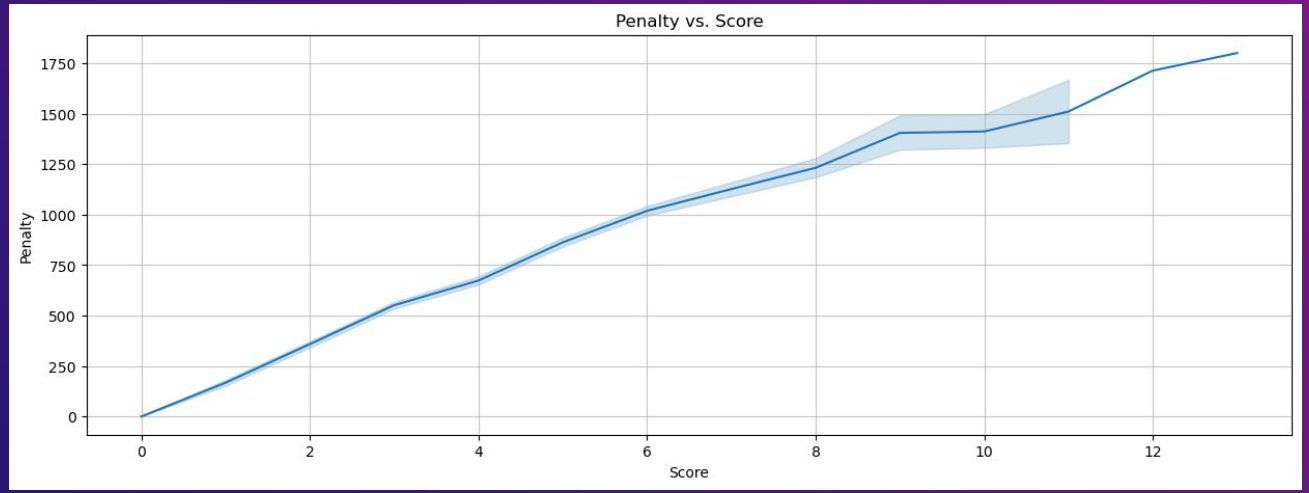
NUMBER OF PARTICIPATING TEAMS



SCORE VS TOTAL PROBLEMS



PENALTY VS SCORE



CONCLUSION

- USA hosts the competition the most with 6 times
- ST. Petersburg National Research University is the most university to win the competition with 5 times (21.7% of years)
- Russia is the most country to win the competition with 15 times (65.2% of years)
- Most participated universities is University of Warsaw and Moscow State University with 23 times
- Most participated team is UCF Knights with 15 times
- In 2018 over 140 team participated the competition and this is the largest number over years
- USA host has the most number of participating teams with over 500 team
- USA Teams are the most participating teams in the competition with over 400 team over years
- There may be a relation between the number of participating teams and the host country or the time the competition is hold
- There may be a relation between Scores and the Total number of problems as from graph best scores are achieved when the total num. of problems is between 12 and 15

04

MODELING & EVALUATION

MODEL USED

- Random forest algorithm with 100 subtrees to predict if the team will achieve a gold medal or no

EVALUATION

- Accuracy score : 0.9927
- Confusion matrix: [[662 1] [4 26]]

THANKS!

IF YOU HAVE FEEDBACK GO AHEAD

amr.mail18@gmail.com
LinkedIn
Kaggle

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon** and infographics & images by **Freepik**