120 Years Olympics Analysis

Group Member: Yuyang Han| Lin Tao

Agenda

- Project Overview
- Problem Setting & Definition
- Data Description & Preprocessing
- Data Exploration
- Data Mining tasks
- Models & Methods
- Performance Evaluation
- Results
- Insights & Impacts
- Q&A

Project Overview

The modern Olympic Games are leading international sporting events featuring summer and winter sports competitions in which thousands of athletes from around the world participate. During the past 120 years, some countries perform better while others perform worse than before. It might be because the overall performance improved or an outstanding athlete emerged.

Project Overview

Predicting the number of Olympic medals for each nation is highly relevant for stakeholders: for example, sports betting companies can determine the odds while sponsors and media companies can allocate their resources to promising teams. Sports managers can evaluate the performance of their teams accordingly.

Michael Phelps

Jenny Thompson

Usain Bolt

Problem Setting & Definition

- Visualize general information about the athletes and the number of medals won by each country
- Hosting Olympics improve performance?
- Predict which countries will be the top of the next Olympics
- Predict the number of medals of each countries at the next Olympics

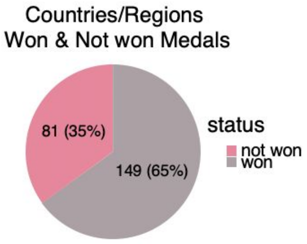
Data Description

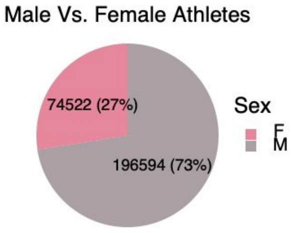
- Data Source 1: Kaggle “120 years of Olympic history: athletes and results”from Athens 1896 to Rio 2016. It contains 271116 rows and 15 columns. Each row corresponds to an individual athlete competing in an individual Olympic event.
- Data Source 2: World Bank, World development indicators. Features including country name, country code, series name, series code, current GDP, GDP growth, GDP per capita, population, population growth and land area.
- Data Source 3: Kaggle “2021 Olympics in Tokyo”, over 11,000 athletes, with 47 disciplines, along with 743 Teams taking part in the 2021(2020) Tokyo Olympics. Only use the medals and athletes datasets.

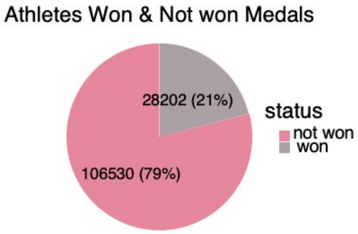
Data Preprocessing

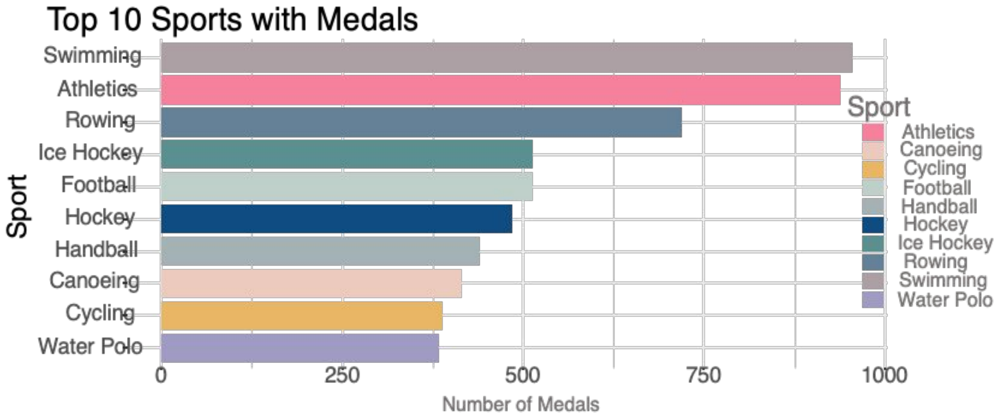
- Remove NAs
- Target Summer Olympics
- Drop Height, Weight, Sex, Age, Event, etc
- Remove duplicate medals
- Find the number of medals for each country in each Olympics
- Add the number of athletes in each country by year
- Add the number of Athletes Normalized, number of Medals during last games
- Merge Country/Region with World Bank Indicators

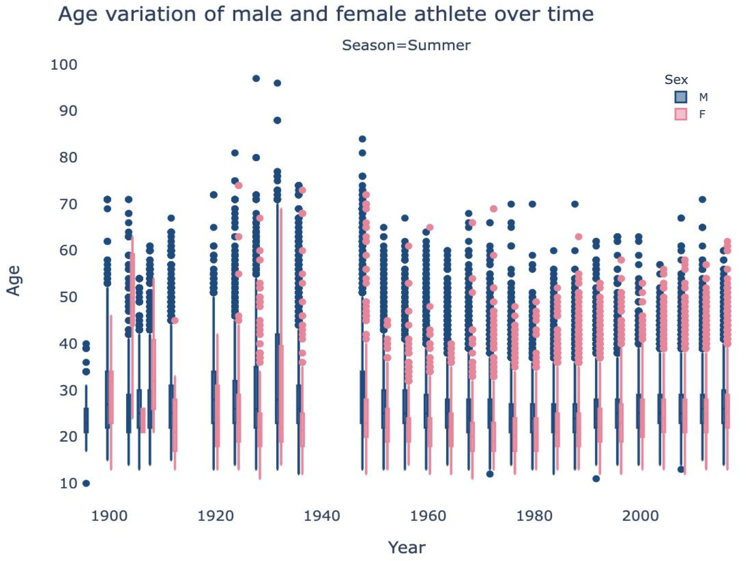
Data Exploration



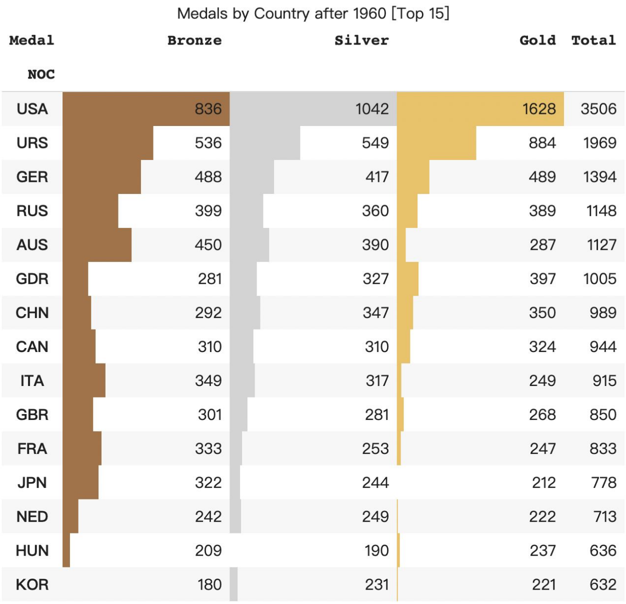






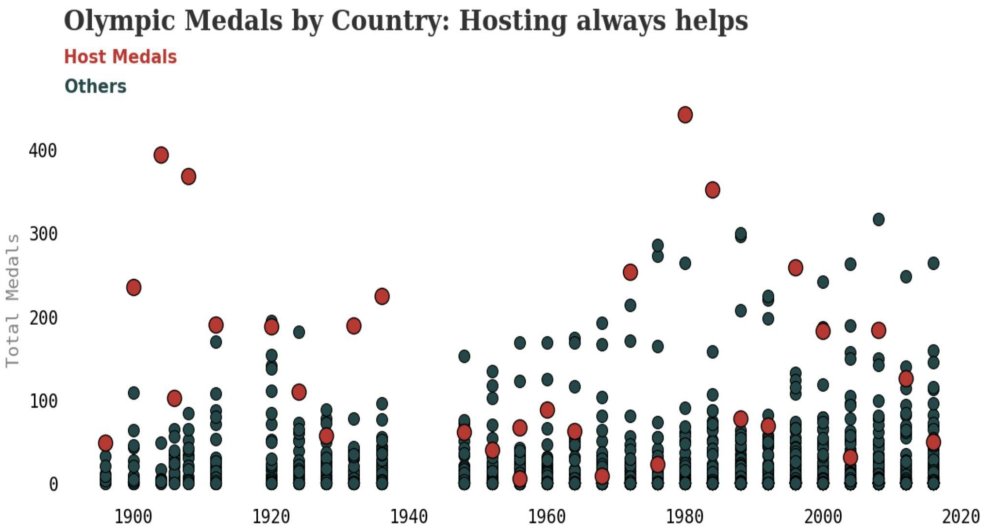


Data Exploration



Medals by Country after 1960 [Top 15]

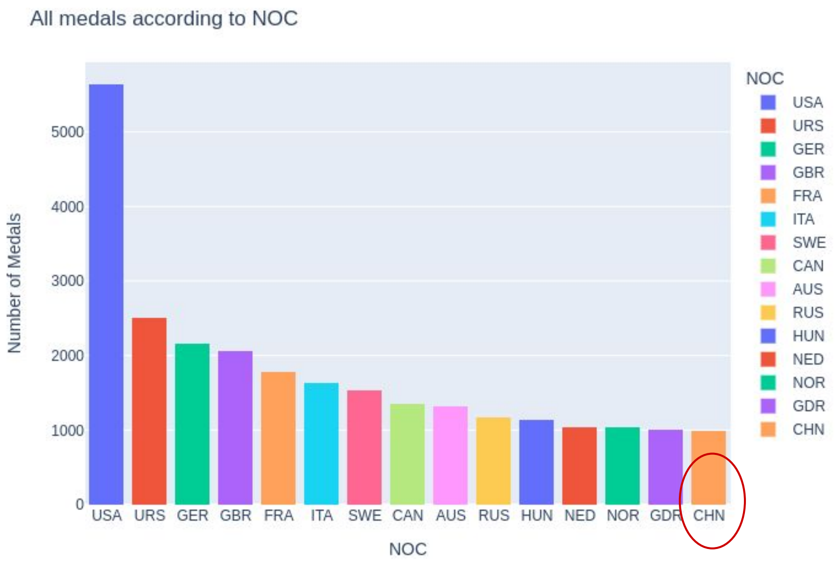
Medal	Bronze	Silver	Gold	Total
NOC				
USA	836	1042	1628	3506
URS	536	549	884	1969
GER	488	417	489	1394
RUS	399	360	389	1148
AUS	450	390	287	1127
GDR	281	327	397	1005
CHN	292	347	350	989
CAN	310	310	324	944
ITA	349	317	249	915
GBR	301	281	268	850
FRA	333	253	247	833
JPN	322	244	212	778
NED	242	249	222	713
HUN	209	190	237	636
KOR	180	231	221	632

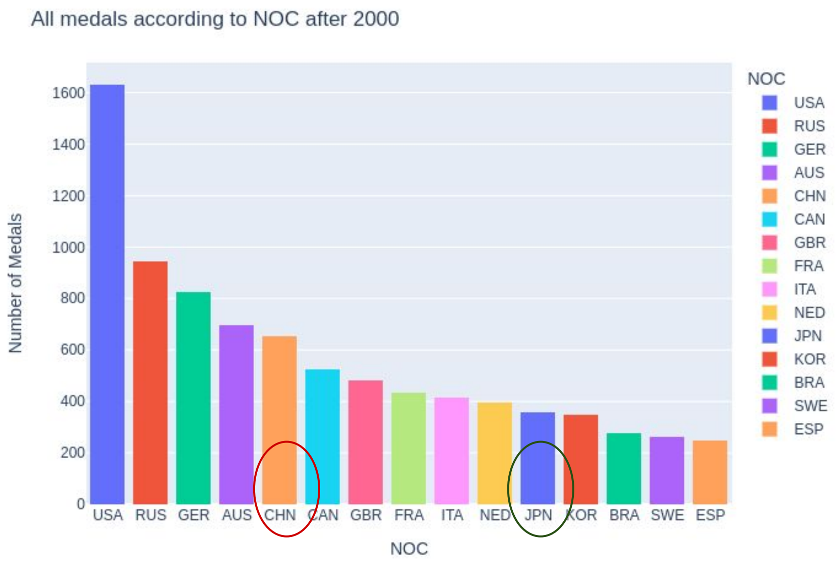


Olympic Medals by Country: Hosting always helps

Data Exploration

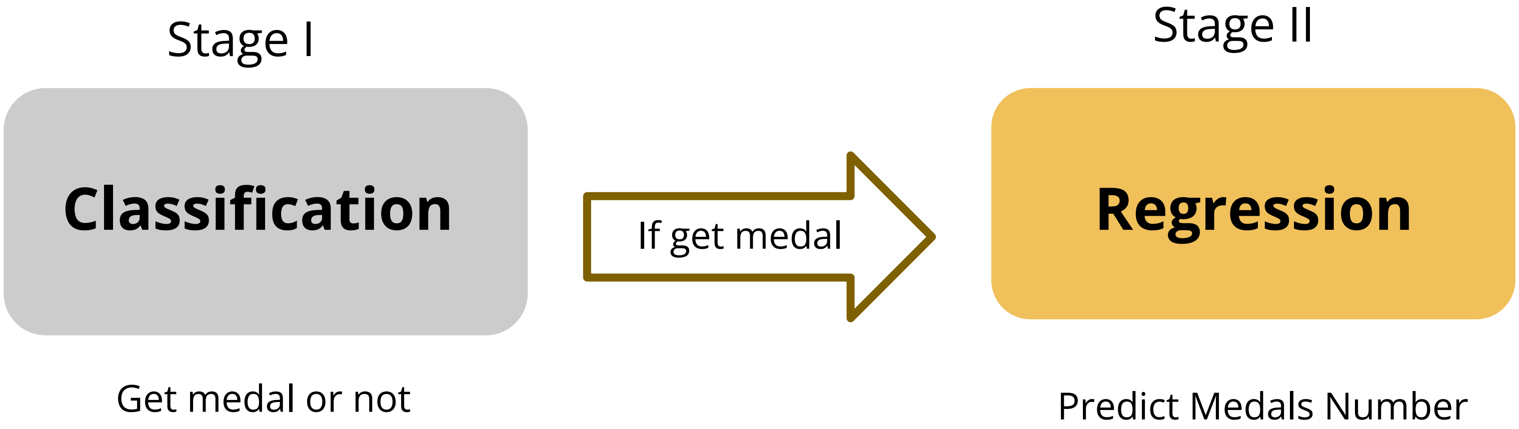
- Some countries perform much better





Data Mining Tasks

- Predict Medal counts of Summer Olympics
- Apply a two-stage algorithm



Stage I

Classification

Get medal or not

If get medal

Stage II

Regression

Predict Medals Number

Models & Methods

Training Dataset: 1980 ~2012; Validation Dataset: 2016

Testing Dataset: 2020

- Classification Model
  - Logistic Regression
  - Gaussian Naive Bayes
  - Random Forest
  - AdaBoost
  - Multilayer Perceptron (MLP)
- Evaluation Metrics
  - Confusion Matrix & Accuracy

- Regression Model
  - Linear Regression
  - Lasso Regression
  - Ridge Regression
  - Poisson Regression
  - Random Forest
- Evaluation Metrics
  - MSE & RMSE

Performance Evaluation

Here are part of our models' evaluation metrics.

- Classification Model
    - Logistic Regression
        - Training Accuracy 0.90
        - Testing Accuracy 0.86
    - Multilayer perceptron
        - Training Accuracy 0.89
        - Testing Accuracy 0.88

- Regression Model
    - Linear Regression (R2: 0.89)
        - Training RMSE 6.75
        - Testing RMSE 5.91
    - Random Forest (R2: 0.93)
        - Training RMSE 5.48
        - Testing RMSE 4.88

Final Choose
Stage I. Logistic Regression –high accuracy and time-saving
Stage II. Random Forest –best performance and could perfectly explain our data

Results

Result of 2020 Olympics Medals Prediction (TOP 5 Countries)

No.	Nation	Prediction	Actual
1	USA	100	113
2	China	85	88
3	UK	77	65
4	Russia	53	71
5	Japan	44	58

Insights & Impacts

- **Sports betting companies** can offer bets on the Olympic medal count for different teams according to our result
- **Media companies** and can reach out in advance to interview and generalize stories about Olympic heroes
- **Sponsors** who profit from signing athletes can get prepared
- **The nation committee** can prepare the Compensation for winning medals; the U.S. Olympic Committee pays $37,500 for a gold medal, $22,500 for a silver, and $15,000 for a bronze.
- **The Government, corporate sponsorship and personal fundraising** can estimate Funds for preparing the team for the next Olympics

Thank You

Q & A