120 Years of the Olympic Games

## Background

*Data*: <https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results>

*Extra data*:

1. World GDP <https://www.kaggle.com/resulcaliskan/world-gdp>
2. 2018 Winter <https://github.com/MrGeislinger/2018-olympics-pyeongchang-data>

*Deadline*: 18 April, 2019

*Presentation time*: 5 mins

*Visualization tools*: Plotly, Dash, D3js

*Useful Tips*:

1. <http://ft.com/vocabulary>
2. <https://www.data-to-viz.com>
3. What’s resampling (a new technique for data preparation)?
4. <https://moderndata.plot.ly/15-python-and-r-charts-with-interactive-controls-buttons-dropdowns-and-sliders/>
5. Logistic Regression: <https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8>
6. Random Forest: <https://www.datacamp.com/community/tutorials/random-forests-classifier-python>

*DS/ML Models*:

1. Random Forest for Regression
2. Naive Bayes for classification to make real time predictions (more suitable for cat data, need to pay attention to variable that is empty/NULL → need to assign value ‘1’ to the empty data for example)
3. DBSCAN (classify cluster by labeling points as ‘noise’/’border’/’core’) for clustering and anomalies, etc
4. ARIMA (Time Series Forecasting)
5. Logistic Regression (sklearn)

*The Grading Scheme for the Mini-Project:*

* 10% for coming up with an *interesting problem* based on the dataset **(DONE)**
* 10% for *exploratory data analysis / visualization* to understand the data **(DONE)**
* 10% for *preparing the dataset* to suit your specific problem definition **(DONE)**
* 20% for the *use of data science / machine learning* to solve the problem **(Time Series Regression, Logistic Regression, Random Forest Classifier)**
* 10% for *learning something new* and/or doing something *beyond the course* **(Time Series Regression (ARIMA), Logistic Regression, Plotly, Random Forest)**
* 20% for the *presentation* of your project, *teamwork*, and overall *impression* **(?)**
* 20% for your *individual contribution*, tested through Q&A after presentation **(?)**

## 

## Timeline

28/3 - 31/3 -> beginning stage, individual parts (clean our data, plotly)

1st of April -> 5pm - 8pm meeting (learning something new)

1 - 7 of April -> Main question = predicting somehow Japan’s performance (only own data)

8/4 - 14/4 -> side questions

15/4 - 18/4 -> presentation & finalization

## Main Question

Performance Prediction at 2020 Summer Olympics

1. **[DONE]** Clean Data

* Convert from float64 to int64 (for weight, height, age)
* Replace NaN with 0
* Get rid of columns “Games”
* Gold, Silver, Bronze = Boolean values

1. **[DONE]** How # of Female / # overall in the team affects the overall performance of a team

* Plotly Timeline (Country Bubbles), X = ratio, Y = # of medals

1. **[DONE]** Predicting age/height/weight distribution among medalists for 2020/2018

* Plotly Timeline: X = age/height/weight, Y = # of medals (with buttons)
* Random Forest = classification (two groups: medal and non medal). Creating a black box with input parameters F(gender, sport, age/height/weight) = probability winning a medal.
* Time Series Forecasting (ARIMA)
* [Summer, Male/Female] = [Athletics, Gymnastics, Swimming]
* [Winter, Male/Female] = [Cross Country Skiing, Alpine Skiing, Speed Skating]

## Other Stuff

1. Exploring data https://www.byrdseed.com/olympic-medal-math-project/

2. Exploring data <https://towardsdatascience.com/to-understand-something-visualise-it-exploring-the-medal-results-of-the-pyeongchang-winter-9cfab4bf0c21>

3. Using of random forest in machine learning?

<https://www.youtube.com/watch?v=eM4uJ6XGnSM>

(To predict outcomes with missing datas)

What factors affects the number of medals(top 3 place) a country can get?(take a few sample of some specific years maybe?)

1. Explore GDP(of each country as a point) vs medals
2. Explore Population Size(of each country as a point) vs medals
3. Explore political stability(of each country as a point) vs medals
4. Average age of athletes from each country vs medals (Categorical?)

For All these points, plot regression line → find anomalies → cleaning of data → find correlation → predict next host for olympic using this?

If not applicable, plot scatter plot for cat graph, gini index, confusion matrix.

Interesting facts while exploring data

1.Height/Weight is Nan for some participants

Countries like Norway do not specify weight due to its policy to combat eating disorder

<https://www.usatoday.com/story/sports/winter-olympics-2018/2018/02/17/norway-policy-athlete-weights/349028002/> (need to maybe put average of height/weight base on their country population data?)

2.Missing values for male athletes in japan in 1940,1944,1948,1916,1920, missing values for male athletes in 1916,1920,1940,1944.(need to maybe put 0 for these years?)

3.Weight correlates well with Height (0.77) for males

4.There some participants with NaN region, some are from the Refugee Olympic Team (10) in 2016,(<https://en.wikipedia.org/wiki/Refugee_Olympic_Team_at_the_2016_Summer_Olympics>),

Team Rika II is a sailing team name for Singapore

Team June Climene is a sailing team name for Singapore in 1960

Unknown Team for art competition at 1916.

Singapore-1,Singapore-2 suspected foreign talent, for table tennis.

5.For Independent Olympians

<https://en.wikipedia.org/wiki/Independent_Olympians_at_the_Olympic_Games>

<https://en.wikipedia.org/wiki/Doping_in_Russia>

Samples

<https://www.kaggle.com/balams/evaluation-of-olympic-games-data-visualization>

<https://www.kaggle.com/tejaeduc/olympics-analysis/notebook>

Japan Data Female Exploration

Winter = [1936…2014]; Missing = [1940, 1944, 1948, 1952, 1956]

Summer = [1928…2016]; Missing = [1940, 1944, 1948, 1980]; Fun Fact: 6 year gap between 92 - 96

To use these factors to predict 2020, maybe to use polynomial regression?

<https://www.youtube.com/watch?v=D9y6dcy0xK8>

<https://images.plot.ly/plotly-documentation/images/python_cheat_sheet.pdf?_ga=2.132745293.2121823054.1555230560-708063730.1553747938>