CFG Project work Document - Group 1

INTRODUCTION:

The Olympics in this day and age consists of all games beginning from 1986 Athens to Rio in 2016. More than just a quadrennial world championship. It provides insight into understanding historical shifts in terms of women's performances and shifting power dynamics geopolitically. Our project focusses on the Olympics from 2008- 2016.

Aims and objectives

In this project, our goal is to show trends over the years in the Olympic history to predict winning trends and use that information to further promote the Olympics and its impact.

Our questions

- How much faster a gold medallist runs in comparison to a silver/ bronze medallist?
- Is there a link between host country and total medals won by host team?
- Does Earning a Medal in an Event mean a Win in Similar Events: Focus on Women's Track events 100, 200, 400, 4 x 100 and 4 x 400 metre Races
- How many events do the top six teams participate in across three years?

In this report, we give a background of our analysis and why we think it can be useful. Then, we go through the steps of analysis: data gathering, preprocessing, visualisation and explanation.

We describe our methods, and the resources we use as well as our results and interpretation of them.

BACKGROUND

Our project is targeted to athletes, coaches and potentially governing bodies who can assist in the promotion of the Olympic games. Our analysis provides detailed insight into specific trends over the years. With machine learning implementation, we could produce models that make these predictions based on potential Olympic participants.

Machine Learning can be implemented to create a model that can be by companies who are looking for sponsorships. This data could be used by companies to single out potential countries to promote their brands. Based on previous wins in specific areas, countries. Brands can use this information to promote business in the top performing Olympic countries.

IMPLEMENTATION AND EXECUTION

Tools and Libraries used: Google Colab, FLASK, Matplotlib, Pandas, Numpy,, Excel to visualise data and then replication in Jupiter Notebooks

Agile Development: there was the use of Google Colab and Meet to review code, Scrum meetings to update on progress and help team members when stuck at any point of the project.

Roles:

Jeanne and Tomiwa: cleaning the data

Nay: creating and implementing the API

Tricia and Abigail: collating and connecting the codes

In general all team members focused on a question and conducted individual data analysis with each team member having specific roles to support as mentioned above

The first step was downloading the dataset from Kaggle using a Kaggle Token. Both a csv and json file were downloaded, after initial inspection of both file types, the csv file was chosen dure to being in a more malleable data

type. Each row of the csv was read in using csv_reader. The data is then turned into a string for transfer over the API server

All data was collected from Kaggle. The main data was the Olympic athletes data , which was cleaned and explored. Null values were dropped from columns and data was reduced to the years we were considering (2008-2016). This exploration was done as a group using Google collab. **https://www.kaggle.com/datasets/vaibhav2025/120-years-of-olympic-history** and this was cleaned through filtering and use of .copy() to pull specific columns that were needed to work on answering this question and saved as datatouse.csv.

The csv file was then cleaned further using filtering by column values to pull the specific track events that was needed for the questions.

Python was used to write the code that loads the CSV into a data frame. The data was then broken down multiple times to address the question before being visualised using of Matplotlib.

Finding an API for Olympic data proved challenging, whilst there were plenty of Olympic databases to choose between, there were no working API's connected to them. To overcome this challenge, this project created an API[1] using FLASK (package for API creation) and ngrok(provides server functionality).

Flask- The API was built with FLASK and contains two test pages. The API server runs through ngrok. The dataset(arg) function in the /<arg> route connects the data we want to send, to the /<arg> page. The API server outputs a http address which is used in [2] for data analysis.

CHALLENGES

A challenge was writing a code that would extract the top 5 countries that participated in the most events grouped by year. As straightforward as it may have seemed, the table was not coming out properly, hence you can see different type of codes I wrote which included resetting the index. In the end I made use of excel which conveniently suggested a way for the data set to be configured to extract the relevant data. I then exported this to a csv file before uploading this on my colab before visualising the final data into a bar chart with multiple bars to represent each year per team. This seemed the most suitable and straightforward in showing the distribution of the event count.

The main challenge of the data analysis and visualisation was pandas DataFrame time dtypes. The time data yielded many inconsistencies in precision and recording protocol. Cleaning the data was attempted in multiple different ways from splitting the data by year to reduce size, then by event for consistency in time precision. Functions such as strptime, datetime and others were used to convert the time variable into a time variable, however when this was achieved, the time could not be analysed. After further bad luck with to_numeric, a function was built to turn the time values, into floats for analysis.

RESULTS

What does the pattern in the amount of events the top 5 teams have participated in across 3 olympics tell us?

Notably from the data gathered, the amount of events the teams has participated in has fluctuated for all 6 teams. The data should be cautiously interpreted as the event count has counted the amount of events each athlete has participated in per team. Most athletes tend to compete in more than one event and as such filtering the data would not have been possible as this is not a duplicate but a variation.

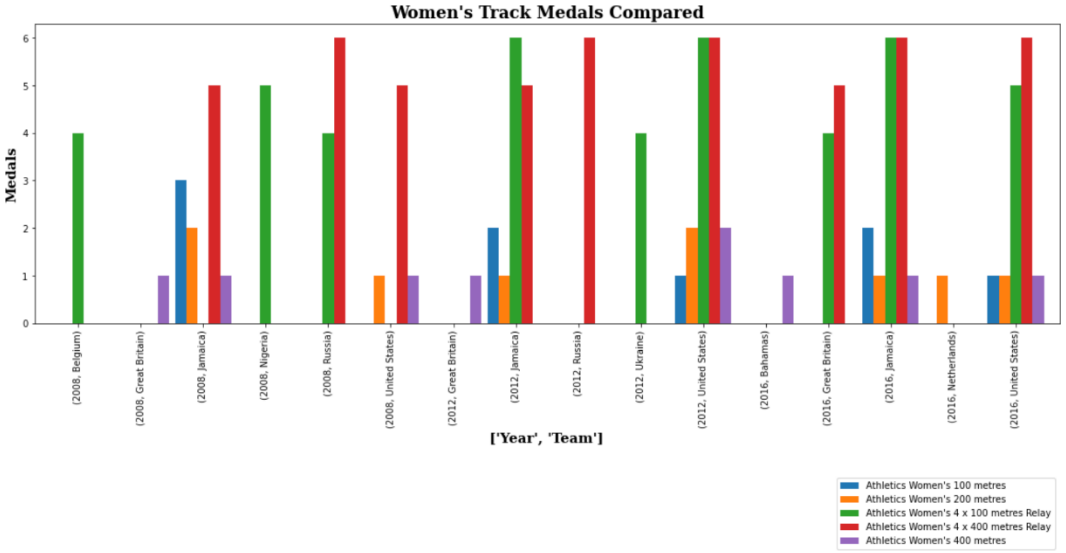
*Side-note; where there is a bar missing, represents the year that the associated country did not make the top 6.

The USA dominates in the amount of events participated in 2008 and 2016 and this could be due to the fact that they have the biggest contingent of Olympic athletes enabling them to specialize in more than one event. Notably in 2012, Britain appeared to be ahead of the U.S., but this is possibly due to the fact that they were hosting the games. Countries hosting the games are told 7 years in advance when they would be hosting. They would therefore have been granted more time and resources to train athletes and enable them to participate in events.

Russia seems to have fallen out of the top 5 countries in 2016. If we give context to the data, this could be due to Russia being provisionally suspended in from all international track and field competitions on 2015 with independent federations being left with the decision-making on allowing Russian athletes to compete. Britain fell out of the top 6 in 2008 however their medal performance was the best in 2008.

To conclude, looking at the event count of each team per year is useful in demonstrating the changes in the amount of events each country participates in and can open up a number of questions relating to whether this can be used as a baseline in determining the success of each team.

Does Earning a Medal in an Event mean a Win in Similar Events: Focus on Women's Track events 100, 200, 400, 4 x 100 and 4 x 400 metre Races



The graph was grouped bar graph to show countries which participated and won at least one medal from 2008 - 2016 in the women's track events that were the focus.

- From the graph Jamaica followed the logic that winning in 100 or 200 metres translated into winning 400metres and the relay race, except the 4 x 100 in 2008 which is blank, it is inconclusive if this was because they didn't participate in the race or if they lost. However judging by the distribution of the number of countries in each year, I think it is safe to assume that it was a loss.
- Team Jamaica (2008 - 11 medals, 2012 - 14 medals and 2016 - 16 medals) and Team United States (2008 - 7 medals, 2012 - 17 medals and 2016 - 14 medals). seem to follow the logic that doing well in one of the races translated into winning in the other similar ones. With consistent and improving performances from the 2008 well into 2016.
- With the analysis, it proves that winning medals in 2 or the track events translated into winning other similar ones, whiles the countries that won in only one kept similar consistent performance of not winning medals in a lot the 4 other races. This applies to the Bahamas, Netherlands and in some years Great Britain.

How Much Faster Does a Gold Medalist Run Compared to Silver/ Bronze Medalist?

This analysis produced three bar graphs displaying the mean time differences between gold-silver medallists, silver-bronze medallists, and gold-bronze medallists.

For the women's 100m(fig.1), the difference between gold and silver medallists is 0.05 S, whereas the mean differences between gold-bronze is 0.12 S; silver-bronze: 0.08 S.

In the 200m(fig2), gold-silver: 0.22 S, gold-bronze: 0.27 S, silver-bronze: 0.05 S. For the 400m (fig3), gold-silver: 0.23 S; gold-bronze: 0.60 S; silver-bronze: 0.37 S.

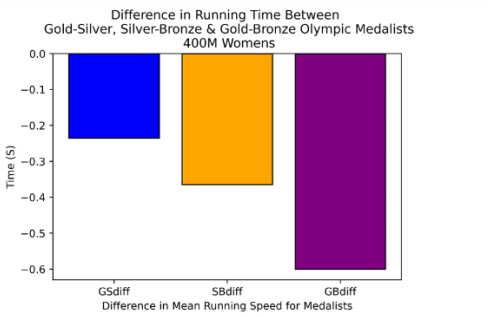


Fig. 3. Bar Chart of Difference in Mean Speed for Gold, Silver & Bronze Olympic Medallists for the 400M Women's Event. Mean difference was calculated as Gold["Time"] – Silver["Time"], thus the results are negative. Gold/Silver differences is plotted in blue, Silver/Bronze in yellow, Gold/Bronze in Purple.

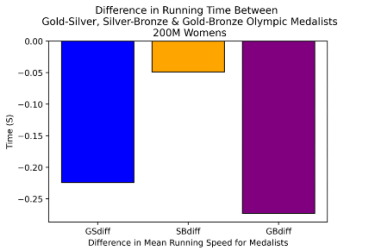


Fig. 2. Bar Chart of Difference in Mean Speed for Gold, Silver & Bronze Olympic Medallists for the 200M Women's Event. Mean difference was calculated as Gold["Time"] – Silver["Time"], thus the results are negative. Gold/Silver differences is plotted in blue, Silver/Bronze in yellow, Gold/Bronze in Purple.

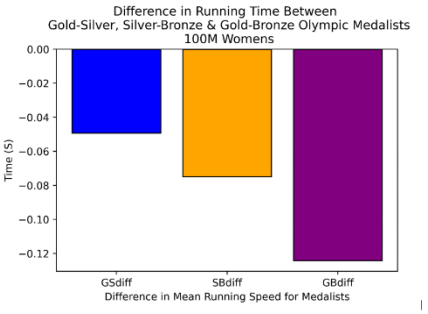


Fig. 1. Bar Chart of Difference in Mean Speed for Gold, Silver & Bronze Olympic Medallists for the 100M Women's Event. Mean difference was calculated as Gold["Time"] – Silver["Time"], thus the results are negative. Gold/Silver differences is plotted in blue, Silver/Bronze in yellow, Gold/Bronze in Purple.

<u>Is there a link between host country and total number of medals won?</u>

From the graph, it is clear that both Team Great Britain and Team China both won the most medals out of the 3 other teams when the Olympics was hosted in their home city. This is a positive correlation.

Team Brazil did not do better than the other 3 teams when the games where hosted in their home country.

Although this was the results in my analysis, there are further things to consider about the findings which may have impacted the results. Such as, total amount of competitors on each team.

The teams may have had more athletes when competing in their home country which gave there a bigger chance of winning more games, as they do not have the potential issue of travelling problems. Or it maybe just, the support of being on their home turf impacting their performance. This is further analysis that could be done at a later date.

CONCLUSION

From the project we can conclude several findings that would be helpful for different stakeholders within the olympic world. Looking at the time difference between the varying medalists would prove to be useful for athletes in implementing the right training to achieve a higher results. The government can consider the correlation between being the host country and the total number of medals won when considering economic policies, particularly investing in the games and sponsoring athletes.

The teams competing in the Olympics can efficiently delegate which athletes perform in each event as the findings show that athletes performing and winning in one olympic event can also attain the same result for a similar event. Lastly the event count calculated allows for a wider view on the the importance for such results and what this means for the top teams dominating the olympics. Whilst there are discrepancies and interpretation that can be skewed by the data, our data has allowed us to gain a bigger insight into wider socio-economic issues and benefits revolving around the Olympics and ultimately allowed us to gain a prediction of future patterns in the teams, athletes and events concerning the Olympics especially on a yearly basis.