Who is *Really* Winning the Olympics?

Kate Osborne *(a1925594)*

Introduction

Once every four years, nearly every country across the globe comes together in competition and celebration of the pinnacle of athletic achievement: the Olympics. It's a tradition tracing back to the first modern Olympics of Athens in 1896, and even further back to the ancient Greek's first Olympiad in 776BC, in honour of their god, Zeus[1]. Whilst the Olympics have changed quite significantly from these early days, now encompassing 32 sports instead of the 6 seen in the beginning, they have also developed a new cultural significance in the broader global landscape.

Today, the Olympics are a platform for more than just athletics, highlighting social issues and inequalities across the world. They are a celebration of every country's unique culture, particular through the host country's performances at the opening and closing ceremonies.

"Citius, Altius, Fortius – Communiter"

The Olympic motto, translated to English, reads: *"Faster, Higher, Stronger – Together"*[2]. It acknowledges the athleticism the games are known for, but also the strong spirit of comradery and mateship displayed at the games.

Despite these strong themes on equality prevalent in the Olympics, there is one area where this is not reflected at all – in fact, it perpetuates current inequalities in society. The medal count often sees the same few countries in the top 5, all of which have large populations[3]. As it only accounts for total medals won, it means small countries who overcome the odds and win a small number of medals will never be able to compete with larger countries, despite the gravity of their achievement in respect to their resources.

This inequality is the motivation behind this investigation. The driving question is "*Is there a fair way to calculate which country is doing the best?*", and thus consequently, "*Who is really winning the Olympics?*". This will be achieved through data analysis and visualisations highlighting areas of inequality.

Methods and Datasets

To complete this investigation, reliable and detailed data was required. Key points of interest were performance in the recent 2024 Paris Olympic Games, population, GDP, number of athletes sent to the games, and location.

To ensure accurate and true medal counts, the data was collected directly from the Olympic website[4]. The data included the number of bronze, silver, and gold medals, as well as the total number of medals won by each country. This data is the backbone of the investigation, as the medal count is the measure of success for a country in the Olympics.

The population data was collected from Worldometer[5], an online site which collates population data from sources such as the United Nations. It contained data across multiple years, but python was utilised to clean it up to just the values desired.

The GDP per capita, purchasing power parity data was collected from the CIA[6] as a csv file. This data

[1] International Olympic Committee. (2024). *Welcome to the Ancient Olympic Games* [online]. Available: https://olympics.com/ioc/ancient-olympic-games

[2] International Olympic Committee. (2024). *Olympic Motto – "Faster, Higher, Stronger – Together"* [online]. Available: https://olympics.com/ioc/olympic-motto

[3] British Broadcasting Corporation. (2021, Aug. 9). *Tokyo Olympics: All the best stats from the 2020 Games* [online]. Available: https://www.bbc.com/sport/olympics/58109921

[4] Olympics. (2024). *Medal Table* [online]. Available: https://olympics.com/en/paris-2024/medals

[5] Worldometer. (2024). *Countries in the world by population (2024)* [online]. Available: https://www.worldometers.info/world-population/population-by-country/

[6] CIA. (2024). *Real GDP per Capita* [online]. Available: https://www.worldometers.info/world-population/population-by-country/

was chosen over the traditional GDP per capita data as it considers how a country's money is distributed, which is truer to a country's financial situation.

To create the model, the factor of most significance (other than medal count) was decided to be population. Each country could be expected to win the proportion of medals that correlates to the proportion of the overall Olympic countries' population that country's population is.

$$P(X) = \frac{Population\ of\ Country}{Sum\ of\ Populations\ of\ All\ Competing\ Countries}$$

Thus, in accordance with the formula, a country which makes up 2% of the total population would be expected to win 2% of the medals available. However, other factors influence a country's ability to achieve success. Financial backing was considered the other major contributor. Whilst specifically the funding a country funnels into sports expenditure would be the ideal measure of financial privilege within athletics, that information was not freely available online in a consistent manner for all competing countries. Thus, an alternative method for measuring finances is gross domestic product per capita, purchasing power parity, which was freely available from the CIA. GDP ranges from $132,000 of Luxembourg down to the $900 of Burundi, making it more difficult to use as a simple multiplicative factor on the population proportion. Thus, it was decided to create a fairer method that does heavily bias countries of low GDP nor high GDP using a logarithmic scale. The natural logarithm of the GDPs ranged from 6.80 to 11.80, thus by dividing that by 10, a multiplicative factor from 0.68 to 1.18 was achieved.

$$Financial\ Factor = \frac{\ln(GDP(PPP))}{10}$$

This will increase the probability of winning each medal for richer countries and decrease it for poorer countries. The product of the population proportion and financial factor yields a probability of winning a medal for each country.

$$P(X) = \frac{Population\ of\ Country\ \times \ln(GDP(PPP))}{10 \times Sum\ of\ Populations\ of\ All\ Competing\ Countries}$$

This probability is for winning each individual medal. By calculating the product of this probability and the total medals available in an Olympics, an expected number of medals can be created. The Paris 2024 Games offered 1039 medals, thus $E(X) = 1039 \times P(X)$.

Using the binomial distribution, the expected medals and real medals achieved by a country can be compared to find the probability that a country earnt the number of medals they did.

$$B(X) = \binom{1039}{medals} \times P(X)^{medals}(1 - P(X))^{\,1039-medals}$$

This will allow comparison of each country – the more statistically improbable the performance, the better a team did. Some countries will perform better than they were expected to, whilst others will perform worse. Their final score will be calculated using the following piecewise function.

$$S(x) = \begin{cases} 0.5 - P(X). & Medals > E(X) \\ -0.5 + P(X). & Medals < E(X) \end{cases}$$

Thus, the countries which exceeded expectations in the most unlikely way will have the highest scores, and those that fell short in the most unlikely way have the lowest scores.

Python was employed to create the dataset and apply the mathematical formulas of the model to each country. The pandas library (as pd) was employed for data cleaning and manipulation. The `pd.merge()` function was used to collate the data sources into one file. This simplified the process, but did require manual editing of data values in country names to ensure all countries matched, as there were some variations in syntax (e.g. *The Gambia* vs *Gambia, The*). The `pd.read_excel()` and `pd.read_csv()` functions were used to import the downloaded data files for their respective file types. The math library was utilised to calculate the sums necessary for the model to work. The `math.log()` and `math.factorial()` functions were used.

Once the data was successfully exported into a csv file, it was accessed in Microsoft Excel to create data visualisations. Microsoft PowerPoint was then utilised to add further visual aids such as flag images or further legend information in a more visually appealing and consumable way.

Experimental Setup and Results

The goals of the visualisations were twofold: firstly, to identify the bias in the total medal tally and medals per capita methods, and secondly, to identify if the new probability model created contains similar biases. Then, if it is decided the new model mitigates the bias sufficiently, it will be used to answer the question of "*Who is really winning the Olympics?*".

The first visualisation is a scatter plot with the population in a logarithmic scale along the x-axis, and total medals won along the y-axis (figure 1). A golden trendline was erected to show the broader relationship between the variables. There is relatively little increase until the 100 million mark, where it sees a sharp, exponential like increase. This visualisation was chosen to demonstrate the bias present in the total medal tally method. A scatter plot was chosen so all 200 countries can be compared visually, and logarithmic scale was used to shorten the graph and make comparison easier.

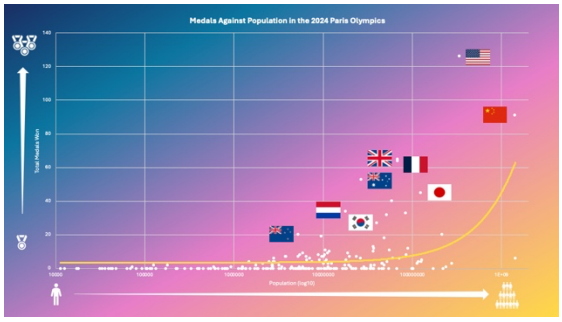


Figure 1. Medals Won in Paris vs Population Scatter Plot

The second visualisation is a tree map demonstrating the number of medals countries would need to equal Dominica's medals per capita (figure 2). The map has a third dimension, where it shifts as larger and larger countries are added which is not demonstratable within a report format but was demonstrated in a presentation format. This data visualisation was chosen as the large difference in area makes the difference between each country highly apparent and does so in a different way to a scatter plot. It also does not require logarithmic scale, which can make the other plots more confusing for those not mathematically minded.

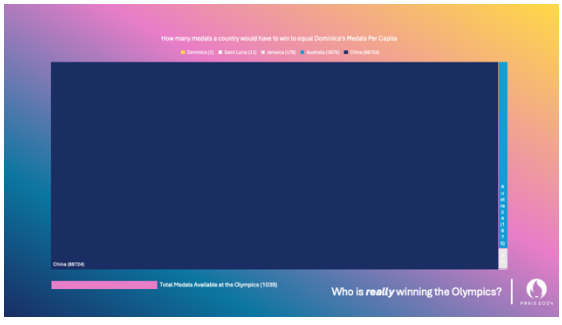


Figure 2. Tree map comparing medals per capita of Dominica against other larger countries

A scatter plot was also created for medals per capita (y-axis) and population (x-axis) (figure 3). This shows on a broader scale than the tree map how population affects medals per capita.

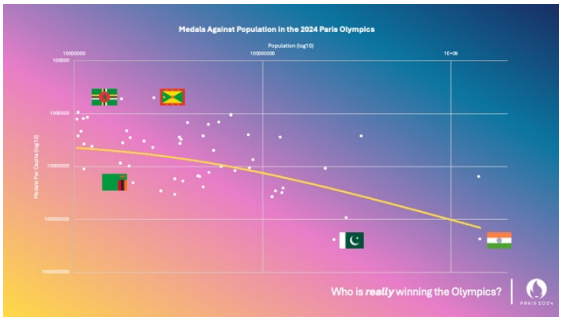


Figure 3. Population vs Medals per Capita Scatter Plot

This column chart is the first visualisation of the new probability model (Figure 4). The purpose of this graph is to demonstrate the range of scores the model can produce. A column graph was chosen to clearly show the separation between the countries who exceeded expectation and those who underperformed.

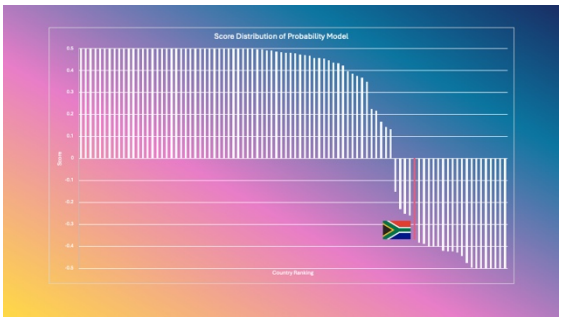


Figure 4. Column graph of probability model score distributions

As with the other medal tally methods, a scatter plot was also plotted comparing the probability model results and population (figure 5). The trendline appears to be much further from the data points in this graph than others, which is corroborated by the low R^2 value of 0.1511 for the trendline.

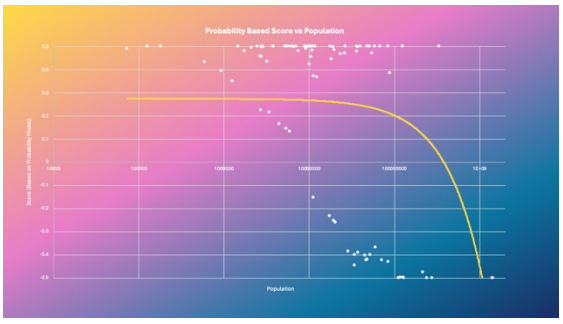


Figure 5. Scatter plot of Probability Model vs Population

The final visualisation is the comparison of all three medal tally methods (figure 6). As they all measure success in vastly different scales, the rankings of

each country regarding different factors were utilised. Each country was ranked 1 – 204 for population, and 1 – 90 for each medal tally method. Countries that did not win any medals were excluded as they create heavily weighted null values that sway trends all in the same way. This visualisation was chosen to be a scatter plot as the trendlines on a scatter plot can clearly show relation and produce R^2 values to compare correlation. The medals per capita had the highest correlation to population, with an R^2 value of 0.5107. The total medals had the second highest correlation, with an R^2 of 0.1347. Finally, the probability model ranking had the lowest correlation, with an R^2 value of 0.0897.

Figure 6. Scatter plot comparing all 3 medal tally methods and their respective relationships with population

These six data visualisations were further analysed to identify trends and possible conclusions.

Discussion and Conclusions

The data visualisations highlighted some clear trends and answered four main questions.

1. Is total medal count biased toward larger countries?

Figure 1 clearly showed this imbalance is true. Whilst the information from figure 6 shows a low R^2 value, the final table did not consider the many smaller countries that did not medal, which influences the medal tally's distribution. Thus, it can be concluded that the medal tally is biased towards larger countries.

2. Is medals per capita biased towards smaller countries?

This hypothesis was also proven to be true. The medals per capita was the method with the most drastic bias, severely favouring small countries, as shown by its R^2 value of 0.5107. Figure 2 also showed this, displaying that countries of a certain size could never equal the figures of smaller countries with the number of medals on offer.

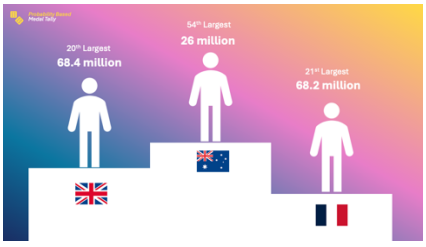
3. Is the probability method biased towards a particular population?

This was the most interesting question out of all the findings of this investigation. The next question, detailing who really won, is irrelevant if the model displays heavy biases akin to the other two models. A true winner should be decided without bias.

The results from the visualisations were positive towards very limited bias in the model. The low R^2 values in figures 5 and 6 showed there was a limited correlation. The spread of high achieving countries across population seen in figure 5 demonstrate that a country can achieve based on this model no matter their size. However, the model does make it incredibly difficult for the extremely large countries, such as India and China, who have populations over 1 billion, to compete. Based on the model, China was expected to win 186.3 medals, which is more than any country has ever won in a single game. Whether it is fair to require such high standards from the country due to their larger population is a question that these visualisations cannot answer. Thus, excluding the high expectations for extremely large countries, the model was seen to have very limited bias in terms of population.

4. Which countries were the true winners of the Paris Olympics?

The countries that had the highest scores in the probability model were Australia, followed by Great Britain, France, Netherlands, and New Zealand. Thus, the model shows that Australia was the true winner of the Olympics. Australia did well as a mid-sized country, coming 3rd overall in gold medals, and having a reasonable total medal tally to keep up with the larger countries. Interestingly, the countries that performed well in either the total medals or medals per capita methods were not in the top 5, showing that their good performance may have just been due to bias towards their population size rather than actual success.



References

British Broadcasting Corporation. (2021, Aug. 9). Tokyo Olympics: All the best stats from the 2020 Games [online]. Available: https://www.bbc.com/sport/olympics/58109921

CIA. (2024). Real GDP per Capita [online]. Available: https://www.worldometers.info/world-population/population-by-country/

International Olympic Committee. (2024). Welcome to the Ancient Olympic Games [online]. Available: https://olympics.com/ioc/ancient-olympic-games

International Olympic Committee. (2024). Olympic Motto – “Faster, Higher, Stronger – Together” [online]. Available: https://olympics.com/ioc/olympic-motto

Olympics. (2024). Medal Table [online]. Available: https://olympics.com/en/paris-2024/medals

Worldometer. (2024). Countries in the world by population (2024) [online]. Available: https://www.worldometers.info/world-population/population-by-country/

Appendix

The feedback I received was minimal, and mostly positive. The positive feedback referred to my idea as 'interesting' (x3) or 'good presentation' (x3).

There were two pieces of feedback which suggested alternative areas of research to explore. The first wanted to find out more about what is more likely to make an athlete succeed. Whilst this would be an interesting topic, to get an accurate and conclusive answer, I would need to do a lot of further collection of data, and it strays outside my research topic. As such, I will not fully explore this aspect.

The second feedback was concerning the participation levels for countries in each sport, with reference to a lack of tropical countries competing in curling. My research only focused upon the summer Olympics, making the curling argument, which is a winter Olympic sport, obsolete. Whilst there is merit to this suggestion, I believe that accounting for too many sources of disadvantage (that are not well outside the realm of sports and thus uncontrollable to athletes) will eliminate much of the variability of results my current model sees. If every possible factor was accounted for, then it would become more and more impossible for countries with slight advantages to be competitive, and too easy for small, disadvantaged countries to win with only a single medal. As such, I will not be adjusting my model to include this factor.

There were only two pieces of feedback that could be classed as constructive criticism. This was slightly disappointing as this is the feedback that is most effective in improving my project, but it is also a positive as it reflects the high quality my project had already achieved.

The first feedback was concerned that my model for calculating the probability of winning each medal only relied upon two statistics, population and GDP. This feedback is in reference to my manipulation of data. I thought this was a good point, and I would have loved to include many more aspects into my graph. However, there were limiting factors such as availability of well documented datasets that are recent enough and consistent for every country competing in the Olympics. Thus, no suitable datasets to add to the model could be found, so no further addition to the model were made. This was not a major concern, as the actual mathematical makeup of the model was not the objective of the assignment.

The second feedback was 'graphs may not be as clear or direct'. This feedback was in reference to my data visualisations. This was trickier to respond to, as it was ambiguous as to why they believed my graphs were not clear. To respond to this feedback, I improved my graphs by adding visual cues as to what each axis represents, as well as including key information in a different colour and using flags to highlight different countries within the diagrams (See figure 1). Whilst it is not clear if this truly responded to what this person could not understand, it attempted to identify any areas tat could cause confusion and thus mitigate that confusion with visual cues.

Name of presenter	Kate
What is good?	Interesting solution to find answers.
What are some risks you can identify?	Graphs may not be clear or as direct.
What might be an interesting aspect to explore?	Can find out more about what makes an athlete likely to succeed.

Name of presenter	Osborne, Kate – Olympics
What is good?	SO GOOD, the maths and the idea is very interesting and yeah overall good work
What are some risks you can identify?	Nothing really
What might be an interesting aspect to explore?	Hmm some winter Olympics sports like Curling for example don't have too many countries competing for it e.g. most tropical countries don't, how does this make it fair for those countries if we're counting medals?

Name of presenter	**Kate Osborne**
What is good?	Cool project, interesting
What are some risks you can identify?	Relying on just two statistics for probability
What might be an interesting aspect to explore?	