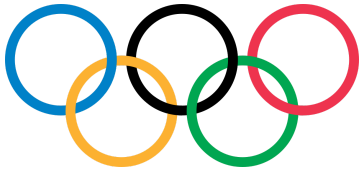
Beijing Wu, Austin Sanders, Melissa McGee
MIDS W200 Summer '21 Section 8
Aug 6, 2021



Project 2: Olympics and Economics

Overview & Motivation:

With the Tokyo 2020 Olympics currently underway, we thought it would be interesting to look at various countries' performance at the Olympics and how it may correlate with various economic indicators.

This paper has three main parts:

1. Overview of our primary Olympics dataset
    - Gold medal in summer games exploration
2. Macro-scale exploration of economic data correlation with Olympic performance
    - Population
    - Birth rate
    - GDP per capita
    - Economic freedom
    - Gender equality (wage gap)
3. Zoomed-in look at individual event domination

1. Olympic Data Overview:

Primary Kaggle Dataset

Our primary dataset is from Kaggle, containing 120 years of Olympic athlete event result data[1]. It contains 271,116 rows and 15 columns. Each row corresponds to an individual athlete competing in an individual Olympic event. The columns provide information on the athlete's name, sex, age, height, and weight; it also provides information on the specific event and outcome that the athlete competed in, including the Olympic game, sport, specific event, and the result. If the athlete medaled, the result denotes a gold, silver, or bronze; if the athlete did not medal, the result is null.

There are 51 unique Olympic games represented in this dataset; the first game is the Summer 1896 games in Athens, Greece and the latest game is the Summer 2016 games in Rio de Janeiro, Brazil. Of those 51 games, there are 29 Summer games and 22 Winter games in the dataset. This dataset reports 13,372 gold medals awarded, 13,295 silver medals awarded, and 13,116 bronze medals awarded.

[1] https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results

An interesting element of this dataset is that for team sports, each individual player is marked as having received a gold medal. We can see this in the example below, where Team USA won Gold in Men's Basketball during the 2000 Summer Olympics.

Name	Age	Team	Games	Event	Medal
Julius Shareef Abdur-Rahim	23.0	United States	2000 Summer	Basketball Men's Basketball	Gold
Walter Ray Allen	25.0	United States	2000 Summer	Basketball Men's Basketball	Gold
Vincent Lamont "Vin" Baker	28.0	United States	2000 Summer	Basketball Men's Basketball	Gold
Vincent Lamar "Vince" Carter	23.0	United States	2000 Summer	Basketball Men's Basketball	Gold
Kevin Maurice Garnett	24.0	United States	2000 Summer	Basketball Men's Basketball	Gold
Timothy Duane "Tim" Hardaway	34.0	United States	2000 Summer	Basketball Men's Basketball	Gold
Allan Wade Houston	29.0	United States	2000 Summer	Basketball Men's Basketball	Gold
Jason Frederick Kidd	27.0	United States	2000 Summer	Basketball Men's Basketball	Gold
Antonio Keithflen McDyess	26.0	United States	2000 Summer	Basketball Men's Basketball	Gold
Alonzo Harding Mourning	30.0	United States	2000 Summer	Basketball Men's Basketball	Gold
Gary Dwayne Payton	32.0	United States	2000 Summer	Basketball Men's Basketball	Gold
Steven Delano "Steve" Smith	31.0	United States	2000 Summer	Basketball Men's Basketball	Gold

This singular win in a team sport resulted in 12 gold medals awarded, rather than just 1 in the dataset. This may then skew Olympic gold medal totals in favor of countries that perform well in team sport events - such as basketball, soccer, 4-person relays in swimming and track - over individual sport events. However, because of various nuances and intricacies associated with collapsing the above information into a single row of data, we made the early data decision to ***not*** collapse team medal information into a single row.

Additionally, the same sports are not consistently present in each Olympic games - the International Olympic Committee (IOC) modifies the sports & events for each Games slightly, "keeping in mind the changing trends, the evolution of existing sports in the itinerary and the popularity of a particular discipline in the host nation."[2] However, due to our limited dataset as is, we made the data decision to ***not*** remove sports/events that are inconsistently present, as that would reduce our dataset too much.

Finally, the dataset contains some null information for some athletes for the attributes `age`, `height`, and `weight`. These variables are irrelevant for our analysis - which uses other columns - so we did not need to clean up these variables in any way.

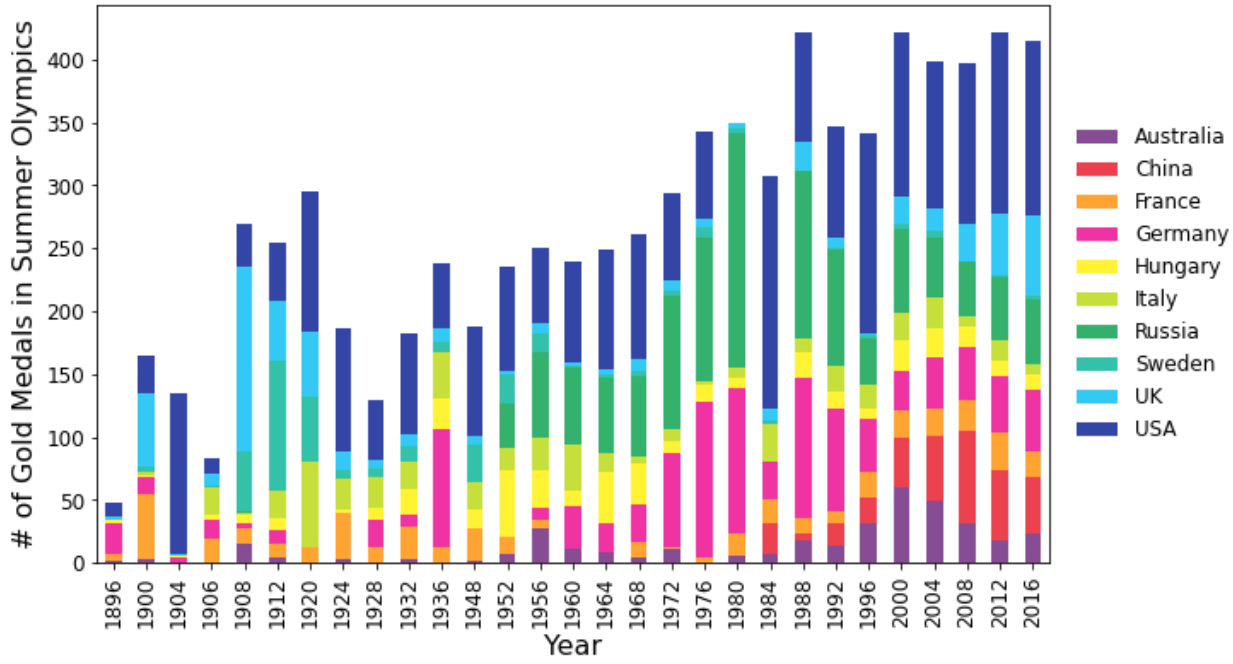
Gold Medals Awarded in the Summer Games

There are two main ways to assess Olympic "performance" of a country: (1) total medals won (gold + silver + bronze), and (2) just gold medals won. Below, we will explore gold medal trends in aggregate. The top ten countries that have won the most gold medals in Summer Olympics are as follows:

[2] https://olympics.com/en/featured-news/how-many-games-sports-tokyo-2020-olympics

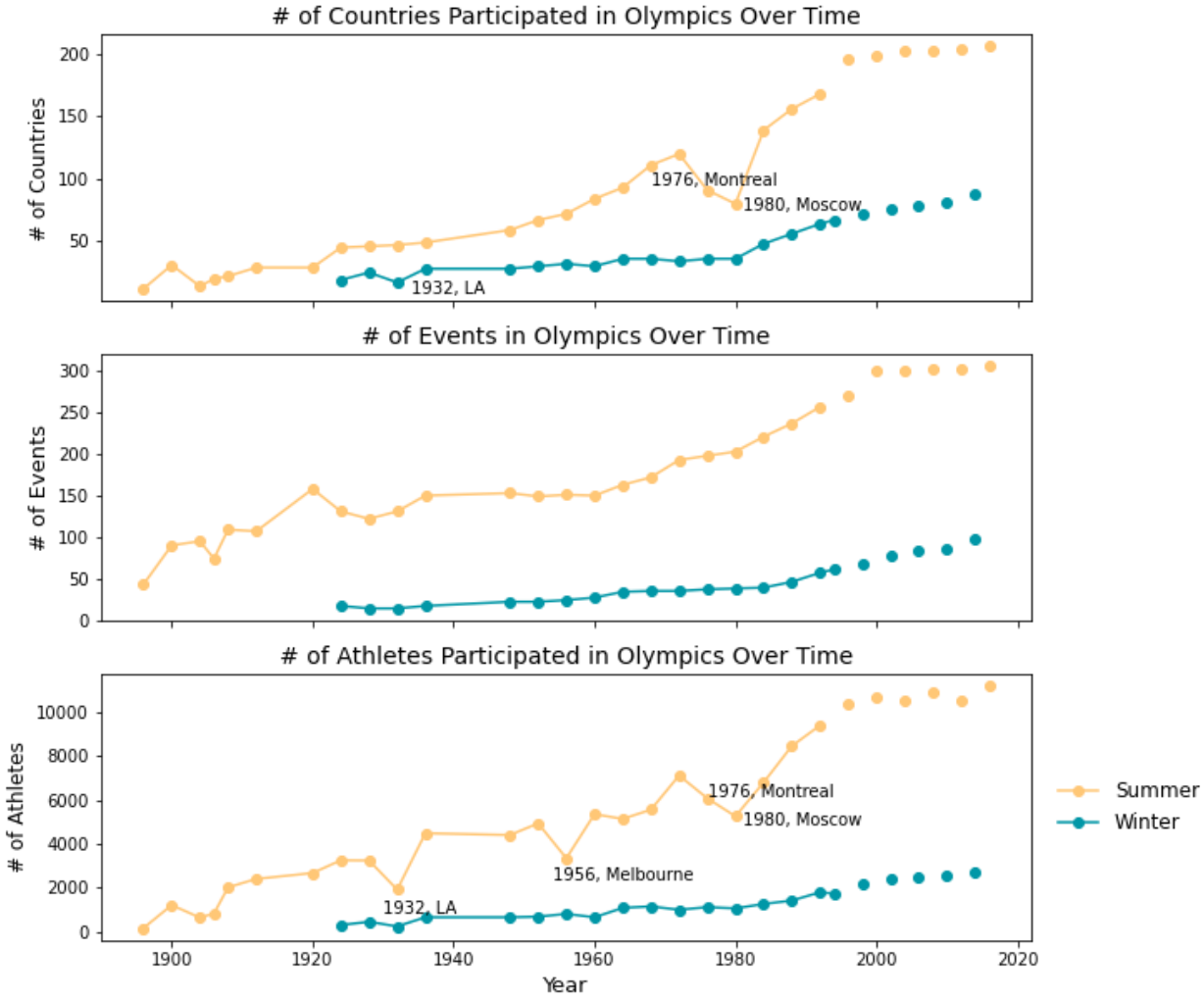
country	medal	counts
USA	Gold	2472
Russia	Gold	1220
Germany	Gold	1075
UK	Gold	636
Italy	Gold	518
France	Gold	465
Hungary	Gold	432
Australia	Gold	362
Sweden	Gold	354
China	Gold	335

Over the years, the total number of gold medals won by these top 10 countries in the Summer Olympics is increasing. An interesting aside is the lack of Team USA gold medals in the 1980 Olympics; Team USA did not compete in the 1980 Olympics held in Moscow as a protest against the late 1979 Soviet invasion of Afghanistan. China has noticeably won a lot more gold medals starting in the 1980s.



This trend - of more medals being won over time - can be explained by the fact that the number of events at the Olympics over time has been increasing and more countries participate in the Olympics over time; hence the number of athletes and medals awarded has been increasing over time. We see that Team USA's boycott of the 1980 Olympics in Moscow affected the number of athletes in those games very clearly. Other notable events (marked in the plots) are the following:

- WW1: Games skipped in 1916
- 1932: Great Depression
- WW2: Games skipped in 1940 and 1944
- 1956: Boycotts due to Israel invading Egypt and Soviets crushing the Hungarian revolution
- 1976: Boycotts in response to apartheid policies in South Africa



Note: Because these plots share axes, Summer & Winter games are no longer "connected" because they are held on alternating 2-year cycles.

2. Economic Data Correlation:

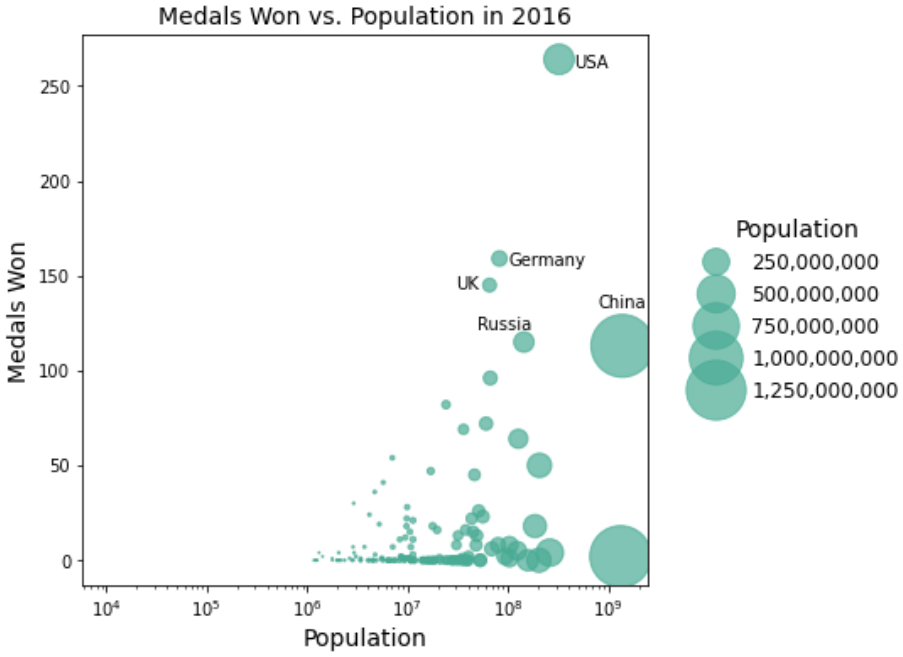
As we can see from the high-level exploration of the Olympic dataset, medal trends are highly dependent on economic and political events such as economic depressions and boycotts. To further explore this, we looked at how various economic indicators - population, birth rate, GDP, economic freedom, and gender equality - may correlate with Olympic performance. Because we are simply combining datasets (not reporting results from a controlled study), we are only able to comment on interesting correlations & trends rather than causation.

For each of these economic indicators, we look at both scatter plots and select countries; the scatter plots allow us to see overall correlations for a specific year, while looking at select countries allow us to see the general trends over time.

Population:

As we saw in the overall Olympic dataset exploration, the number of gold medals awarded each Olympic games increases over time, due to additional countries participating, additional events being included, and additional athletes participating in the Olympics over time. Another way to understand medal trends over time is to understand the correlation between a country's population in a given year and the number of medals it wins during the Olympics that year.

To this end, we utilized a dataset[3] from the World Bank that provides the total population for each country over time. There are 266 unique countries in this dataset, spanning from 1960 to 2020. We joined this dataset onto our primary dataset so that we could explore interesting relationships between country populations and Olympic performance.



Another interesting way to look at correlation between population and Olympic success is looking at a specific country, and how the medals won over time may change with population. We can see here that it does not seem as though there is any correlation between Olympic performance and the country's population.

[3] https://data.worldbank.org/indicator/SP.POP.TOTL

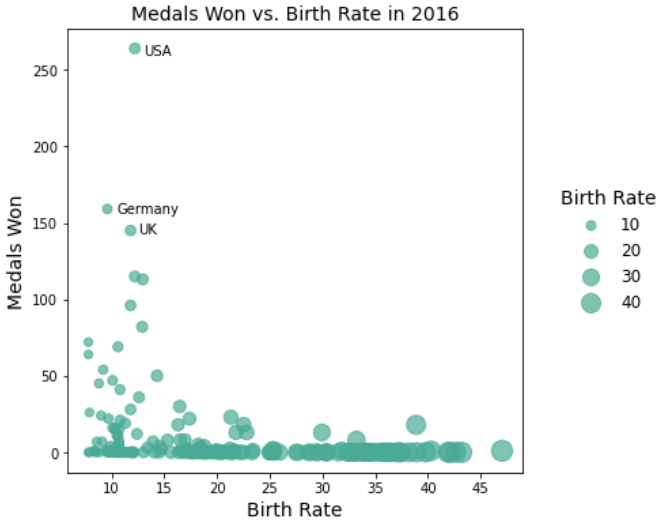
We can see here that these time series plots seem to confirm our conclusion from the scatter plots that there doesn't seem to be a visually discernible correlation between a country's population and their performance at the Olympics.

Birth Rate:

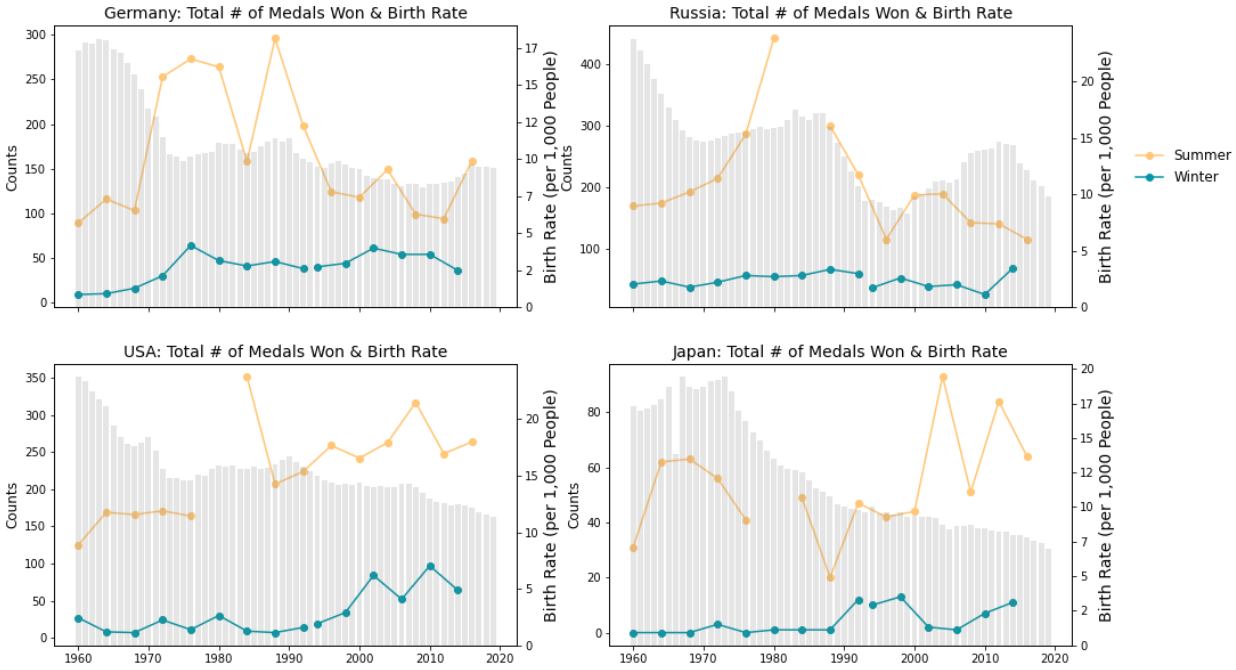
While we see that population seemingly does not correlate with Olympic performance, that may be because the overall population encompasses many different other economic indicators - birth rate, death rate, immigration, emigration, etc. Thus, we found it motivating to look at specifically the birth rate, as studies have shown that there exists a "correlation between lower birth rates and economic distress"[4].

To this end, we utilized a dataset[5] from the World Bank that provides the birth rate for each country over time. There are 266 unique countries in this dataset, spanning from 1960 to 2020. We joined this dataset onto our primary dataset so that we could explore interesting relationships between country birth rates and Olympic performance.

[4] https://www.pewresearch.org/social-trends/2011/10/12/in-a-down-economy-fewer-births/
[5] https://data.worldbank.org/indicator/SP.DYN.CBRT.IN?view=chart



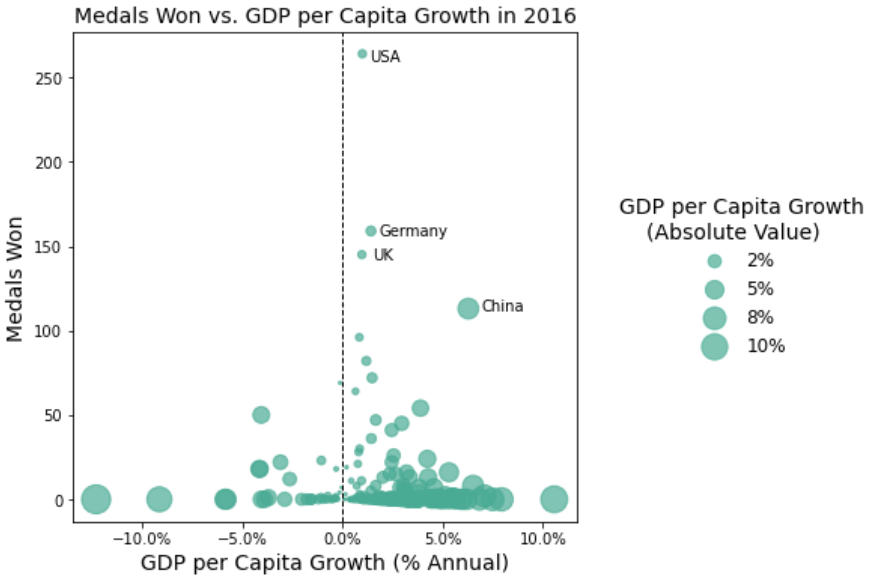
The scatter plot does not seem to indicate any correlation between the birth rate of a country and the medals won for that country in a given year. However, there may be trends worth discussing if we look at the data over time:



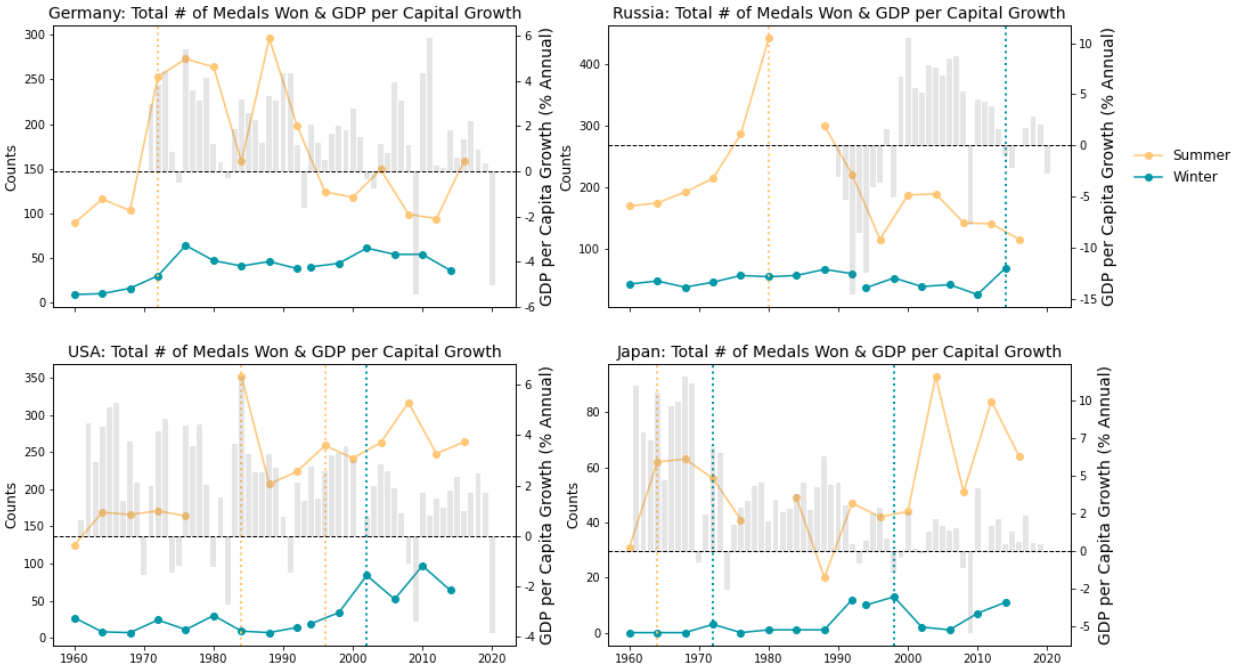
We see, interestingly, that Olympic performance seems to lag behind birth rate, though this may be coincidental with general higher medal counts in the 1980's and beyond. These select countries below had a "baby boom" in the early 1960's; those babies would then be around competition age (18-24 years old) in the 1980's. This may help explain the ~20-year lag in birth rate and Olympic performance. Likely, any trends seen here are entirely coincidental, though it *is* interesting / entertaining to wonder if baby booms impact Olympic performance decades later.

GDP Per Capita:

The implicit link between birth rate and Olympic performance is that birth rate is indicative of economic performance, often indicated by gross domestic product (GDP). We thus thought a similar look at Olympic performance as it relates to GDP would likely yield the highest correlation. To this end, we utilized a dataset[6] from the World Bank that reports the GDP per capita of each country over time (in current US dollars).



Again, the scatter plot does not seem to indicate any kind of correlation, but again the scatter plots are limited to looking at a snapshot in time for countries. Thus, looking at individual countries over time may be more illuminating:



[6] https://data.worldbank.org/indicator/NY.GDP.PCAP.KD.ZG?view=chart

We can see above the clearest example of correlation yet. There is clearly a correlation between a country's GDP per capita and the total medals won by that country. Indeed, there may be a confounding variable that is the closer link - a higher GDP is correlated with a country's ability to send more athletes to the games, and the more athletes are sent to the games, the higher chances of medals being awarded. No matter the exact linkage, the correlation above is very clear.

This plot also shows vertical dotted lines to denote when those countries hosted the Olympics. We can see that while host years do not seem to have a correlation with GDP per capita growth, it does provide a boost in the total number of medals won.

Freedom Index:

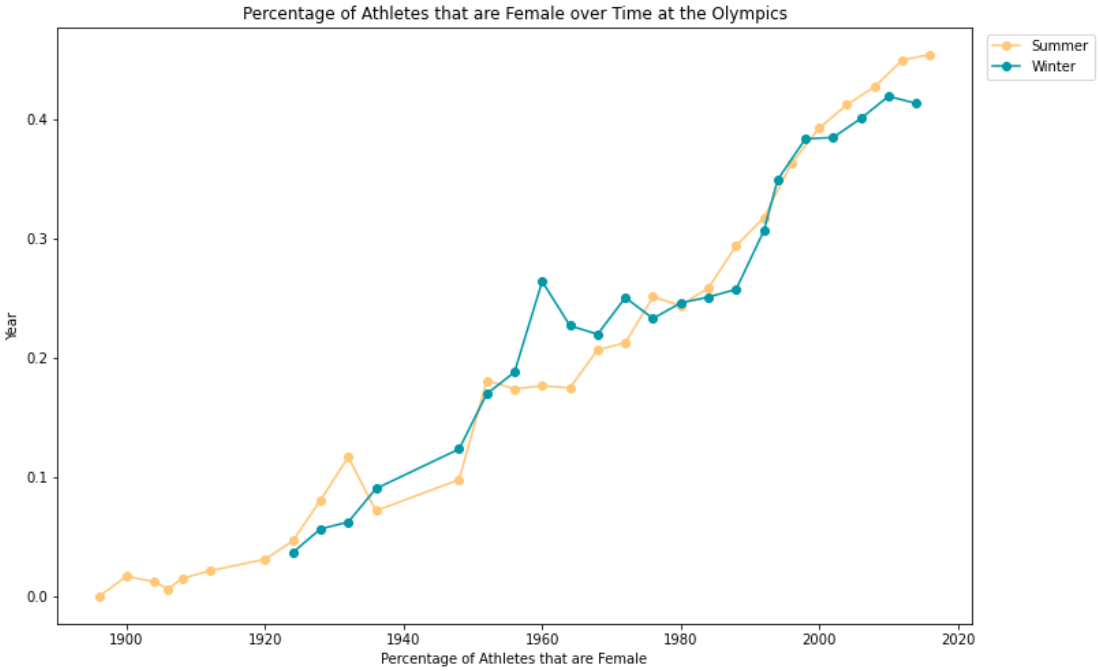
Countries ranking in the top 20 globally for the 2016 Heritage Freedom Index totalled 859 medals at the 2016 Olympics in Rio. The bottom 20 countries in the Freedom Index ranking totaled only 72 medals at the same games. War-torn states (Afghanistan, Libya, Iraq, Syria, Sudan) that did not receive a Freedom Index did not win any medals at the 2016 Summer Games.

Team	Total Medals	Freedom Score
Hong Kong SAR	0	88.55
Singapore	1	87.78
New Zealand	36	81.56
Switzerland	11	81.03
Australia	82	80.34
Canada	69	77.97
Chile	0	77.66
Ireland	3	77.31
Estonia	4	77.22
United Kingdom	145	76.41
United States	256	75.44
Denmark	41	75.26
Lithuania	7	75.23
Taiwan	0	74.74
Mauritius	0	74.73
Netherlands	45	74.55
Germany	157	74.37
Bahrain	2	74.25
Luxembourg	0	73.86
Iceland	0	73.34

Team	Total Medals	Freedom Score
Korea, North	0	2.30
Cuba	11	29.79
Venezuela	3	33.74
Zimbabwe	0	38.23
Turkmenistan	0	41.89
Eritrea	0	42.70
Congo, Republic of	0	42.80
Iran	8	43.49
Equatorial Guinea	0	43.67
Argentina	22	43.77
Central African Republic	0	45.23
Timor-Leste	0	45.76
Uzbekistan	13	46.02
Kiribati	0	46.22
Chad	0	46.33
Congo, Democratic Republic of the Congo	0	46.38
Ukraine	15	46.76
Solomon Islands	0	46.98
Bolivia	0	47.40
Ecuador	0	48.56

Gender Equality (Wage Gap):

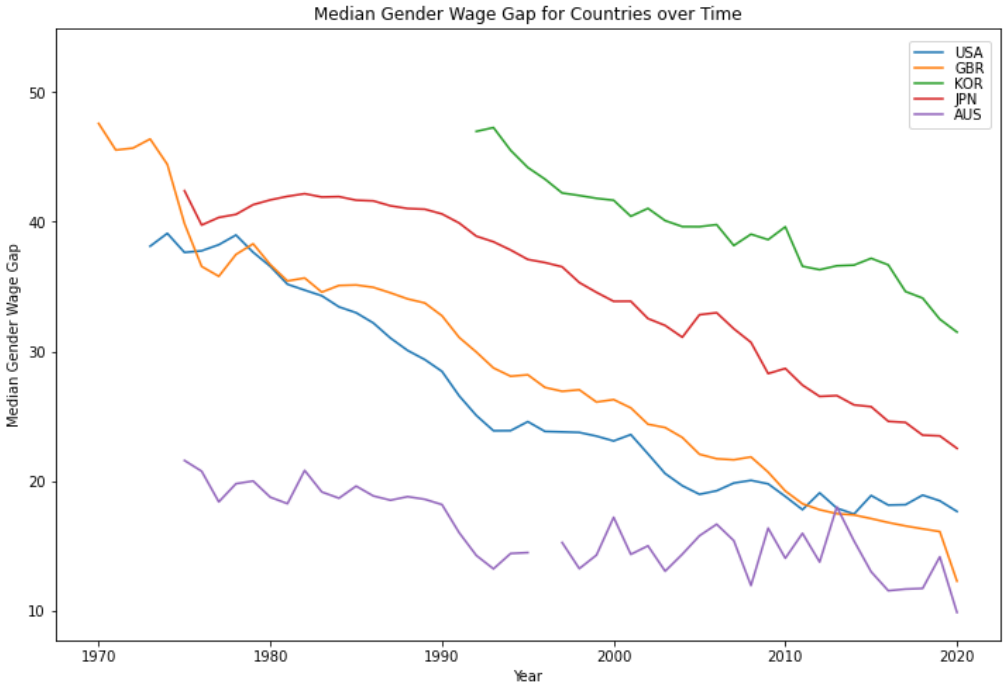
The Olympics have come a long way since the inaugural 1896 Olympics in Athens, Greece, where women were not allowed to compete. While still not fully equal, 45.5% of the athletes at the 2016 Olympic games in Rio de Janeiro were women, an all-time high.



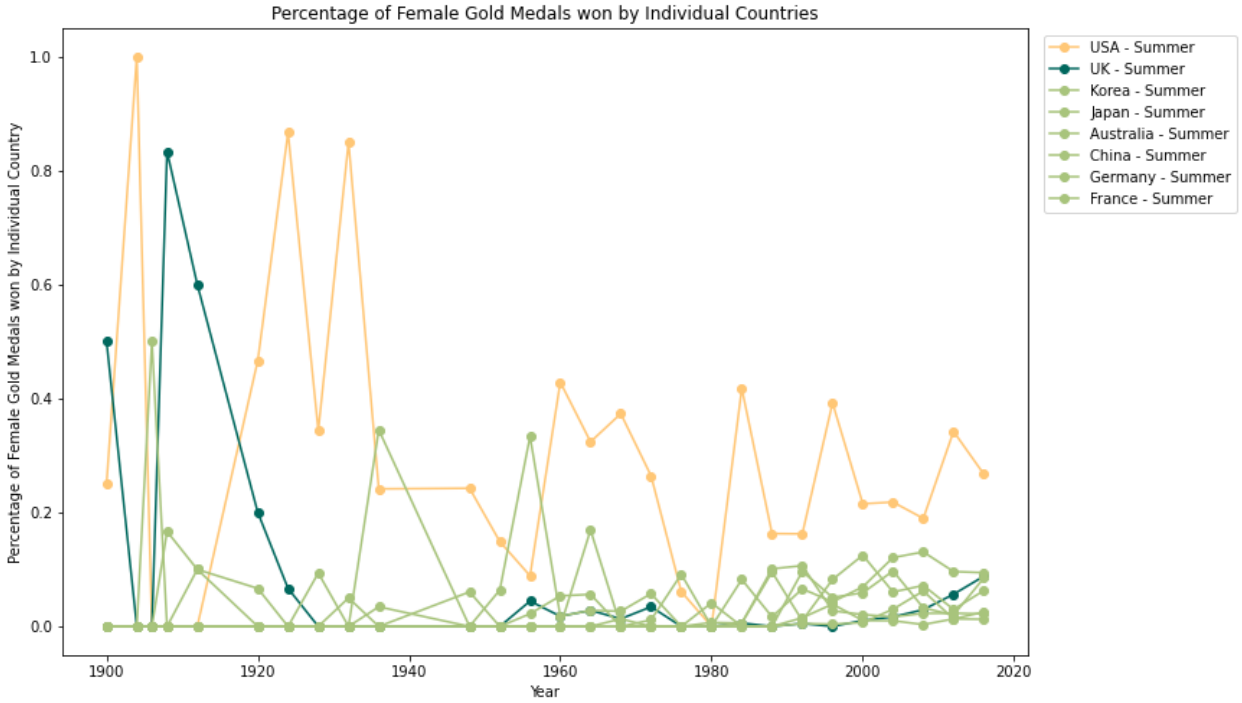
The data above is on a macro-scale, and it would be motivating to look at this on a per-country basis, especially as different countries have different levels of gender equality. There are many different ways of assessing gender equality, but one that is very indicative is the gender wage gap. The higher the gap, the worse the gender equality in that country.

To this end, we utilized a dataset[7] from the Organisation for Economic Co-operation and Development (OECD) which reports the gender wage gap as “the difference between median earnings of men and women relative to median earnings of men”. This dataset is far from perfect - there are data starting from 1970 at the very earliest for Great Britain, but many countries have much less data. As such, five countries with the largest amount of data, and the greatest diversity in gender pay gap percentages, were chosen to display and investigate further: United States (USA), Great Britain (GBR), Korea (KOR), Japan (JPN), and Australia (AUS).

[7] https://data.oecd.org/earnwage/gender-wage-gap.htm

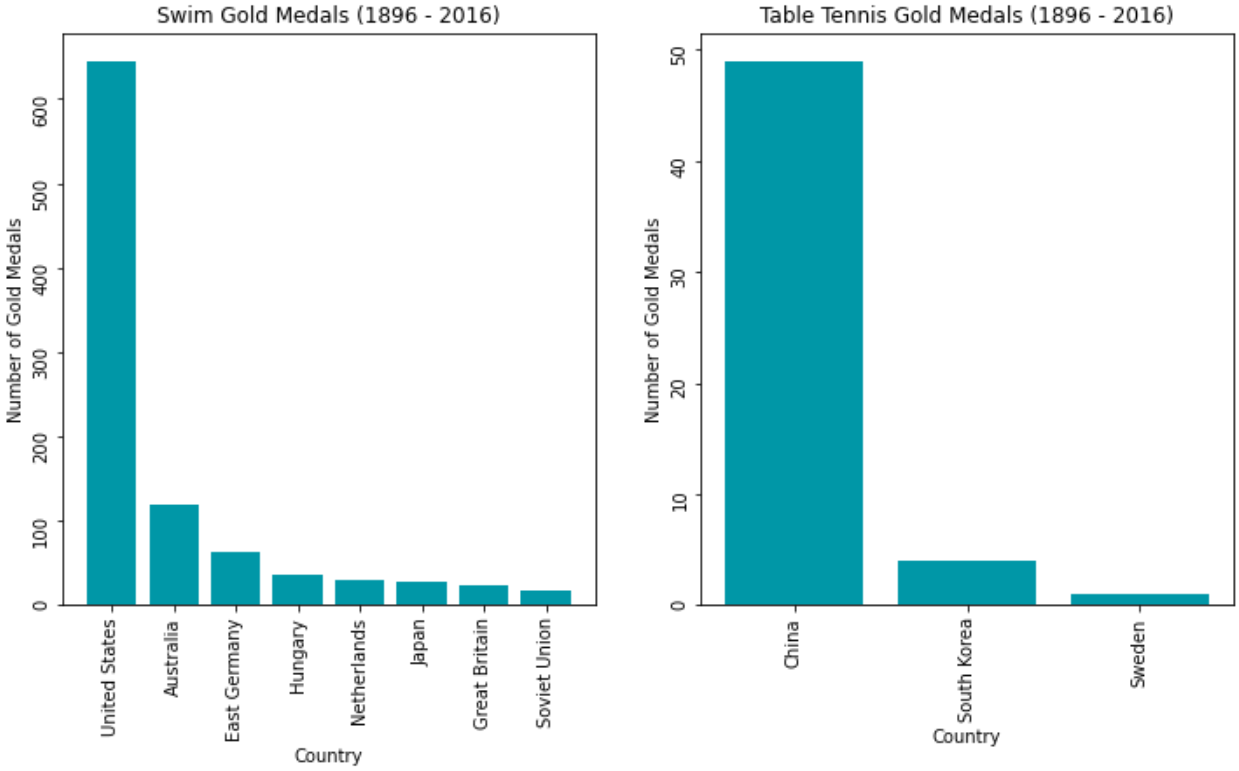


This plot above then raises the question of whether countries with lower wage gaps - such as Australia - have women that tend to perform better compared to women in countries with higher wage gaps - such as Korea. However, when looking at female performance at the Olympics by looking at the gold medals won by women in various countries as a percentage of all gold medals won by females, the results tell a different tale. While Team USA continues to dominate (as it does generally at the Olympics), there doesn't seem to be a correlation between countries with higher or lower wage gaps and the performance of their women at the Olympics. Instead, the trends seem to imply as gender equality increases globally, there is greater diversity in the countries where women tend to perform well.

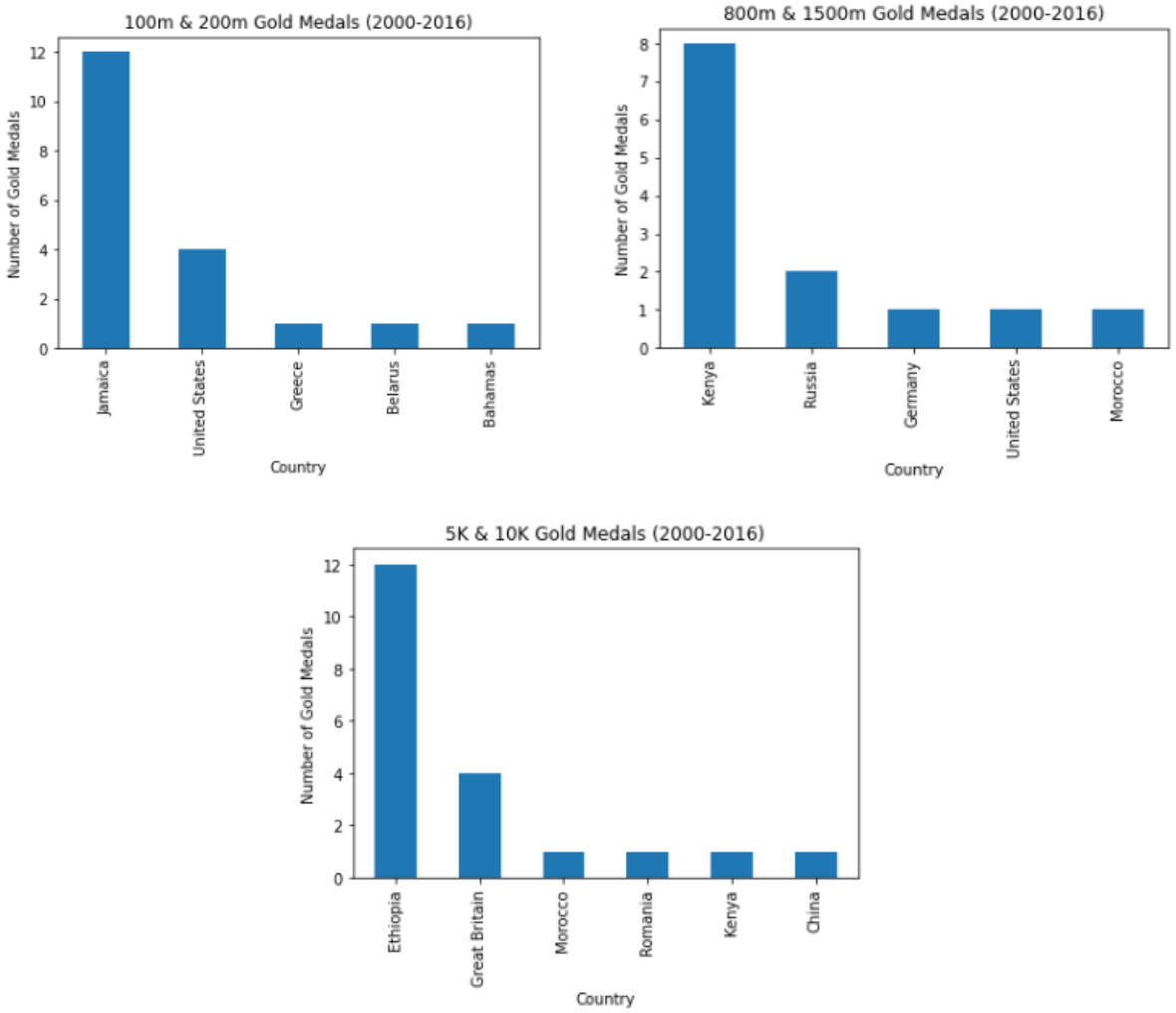


3. Individual Event Performance:

While it is certainly interesting to look at country performance & economic indicators on a macro scale, it is useful to “zoom in” and look at sports where there is a clear domination by a certain country. We see that the United States dominates at swimming, earning over 600 gold medals in the sport. We also see China dominates at table tennis, winning a vast majority of most gold medals ever awarded in the sport.



We identified an interesting change in Athletics (Track & Field) performance for individual running events. The United States of America won 45 out of 132 gold medals in the 100m, 200m, 800m, 1500m, 5000m, and 10000m events (male and female) over the first 100 years of the games. However, in the Sydney 2000 Summer Games, new countries began to emerge on the scene, demonstrating unparalleled success in the running events.



While a triumphant two decades on the track, Jamaica, Kenya, and Ethiopia cannot breakthrough in any other sports. All of their gold medals, except one, in the history of the Olympics have come in Athletics (Track & Field).

Sport	Kenyan Gold Medals	Jamaican Gold Medals	Ethiopoian Gold Medals
Athletics	33	38	22
Other Sports	1	0	0

Concluding Thoughts & Ideas for Further Exploration:

In all, we found the following correlations between Olympic performance and economic indicators:

- Population: none
- Birth rate: A potentially coincidental 20-year lag correlation
- GDP: Seemingly correlated, especially when the country hosts
- Freedom index: More “free” countries perform better in aggregate than less “free” counties

- Gender equality (wage gap): General trends reflected in increasing diversity of women who win gold

During this data exploration, we thought of a few interesting ideas for further analysis:

First, it would be interesting to examine the economic impact of hosting the Olympics. While there have only been 51 games (and hence 51 host cities), anecdotal evidence provides mixed reviews on the benefits of hosting the Olympics. While it does attract a multitude of international visitors, the host countries also typically spend billions building facilities that may never see a second use again.

Additionally, the Olympics dataset was massive, and we never looked closely at the individual athletes. USA athlete Eddy Alvarez is set to be the first individual to win an Olympic medal at the Winter and Summer Olympics. Michael Phelps won 28 total medals over his Olympic career. We would like to identify trends in what makes these athletes successful. It would be interesting to find which countries produce athletes that compete over the course of multiple Olympic Games and dominate in their respective event. We could see which sports consistently have one generational athlete winning the most gold medals.