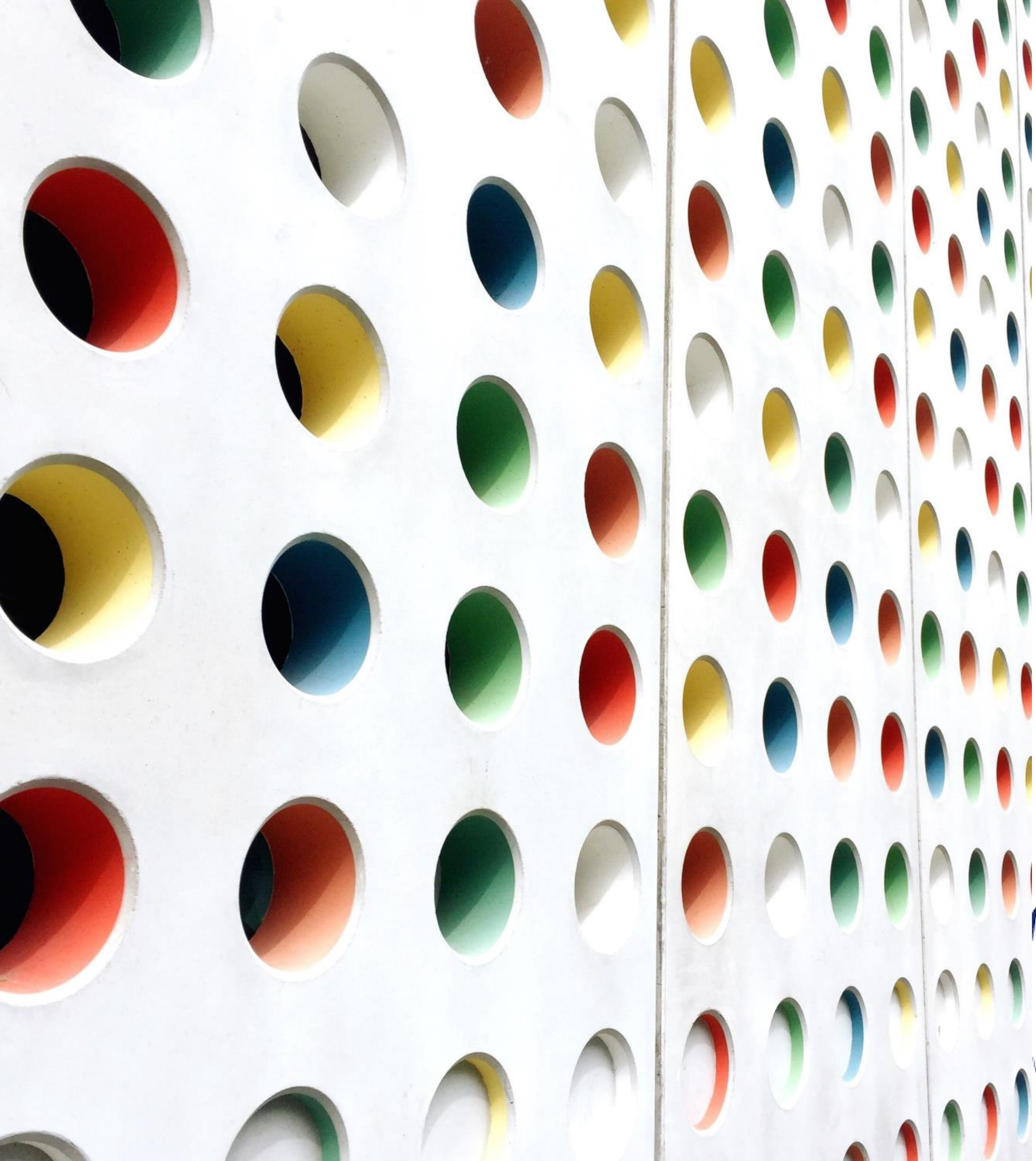


Analysis of the Olympics:

Initial findings for The Sports History Group at Swansea University

T.H.SIMM

Swansea University
Prifysgol Abertawe

Background & Data

Athletes

Countries

Games

Background & Scope

- This report follows an initial meeting with The Sports History Group at Swansea University

- The group are interested in various aspects of the Olympics, how it reflects history and changes in athletes

- Based on an initial meeting a broad methodology was established:

◦ A database of Olympic data was provided

Three broad questions were set to be investigated

1. What are the characteristics of athletes? How does this change with time, and can it be linked with societal or global changes?
2. What countries do better at the Olympics? Is there a way to quantify this?
3. What is the influence of a games being a home event?

Interactive Presentation

Ask questions during the presentation

Data

The Data

Two csv files (representing two different tables)

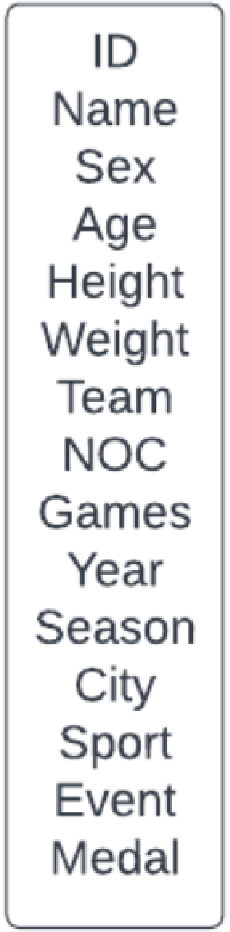
1. athlete_events

- Represents the athletes competing in the Olympics
  - Details of the athlete:
    - Name, Sex, Age, Height, Weight, Medal Won
  - Details of who they represent:
    - Team, NOC (both Team and NOC represent country)
  - Details of the game attended
    - Games (name of games), Year, Season (Summer/Winter), City
  - Details of the event they participated in
    - Sport, Event (event is a subcategory of sport)
- The main table consists of 270,000 rows, with ~135,000 unique names in the table

2. NOC

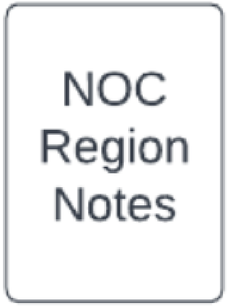
- Additional information about the countries

athlete_events



ID
Name
Sex
Age
Height
Weight
Team
NOC
Games
Year
Season
City
Sport
Event
Medal

NOC



NOC
Region
Notes

The Data

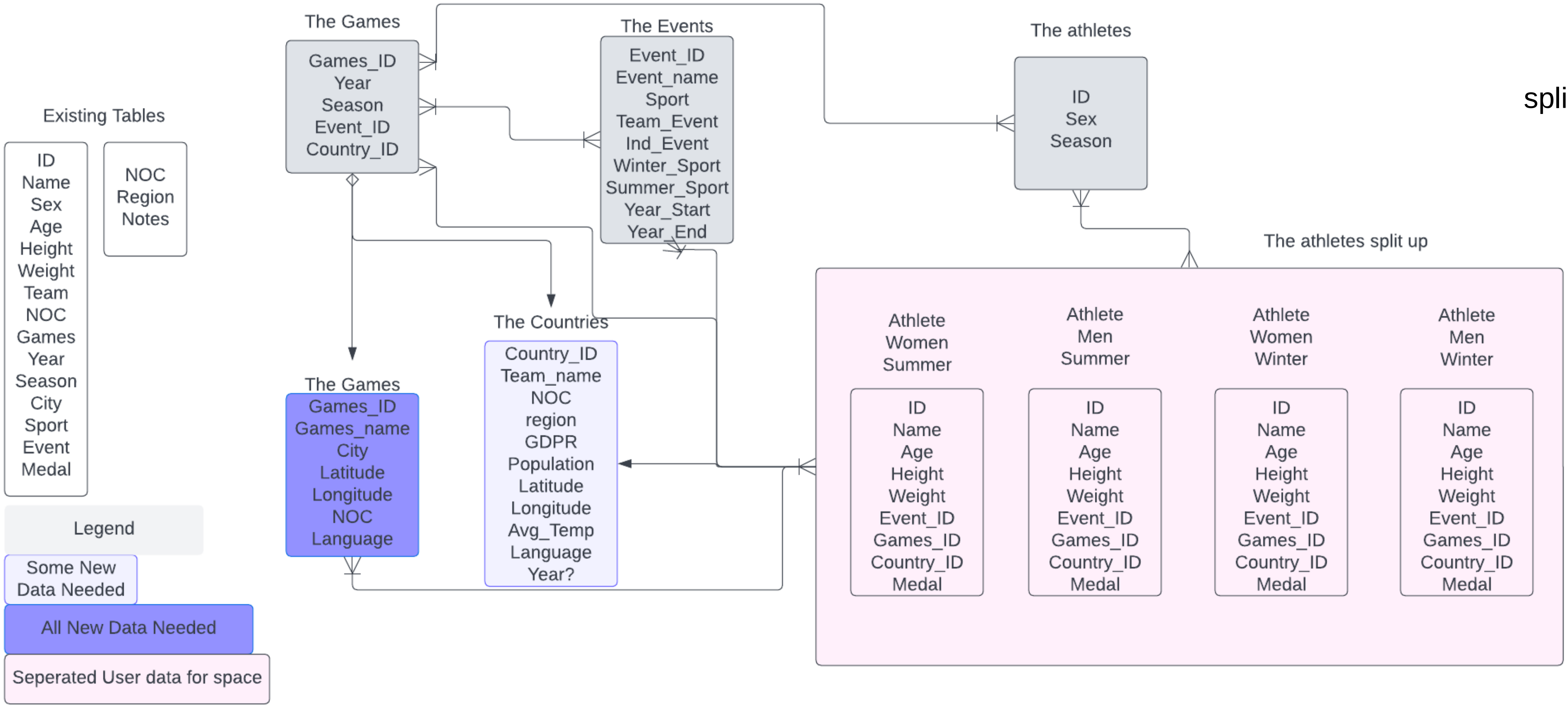
- Lots of columns and lots that are objects (i.e. strings)
    - so, we want to refine this by reducing columns and making it an integer or something smaller than object if possible
- There are some NaN values, particularly for height/weight at earlier games and also for medals
- An athlete can be represented in several rows if they do multiple events or at different games (e.g. Christine Jacoba Aaftink). So we may want a separate ID that incorporate the athlete and the event/games that is unique
- The TEAM, NOC we only want one identifier and a separate table for countries

Steps taken

- First the users are split up based on whether they are male or female and whether they are in the summer or winter games. So split into 4.
- Secondly not all data is needed for these athletes table, so instead of 15 columns this is reduced to 9
- Thirdly, the size of these athlete table is reduced by replacing several variables from string to int to reduce the size. Since for example, there is only a limited number of events.

The Data

An entity relationship diagram (ERD) of the tables described above was developed as shown below.
Lucid Chart was used to produce the ERD.
There may be too much splitting here which would be considered for the next iteration



Existing Tables

ID
Name
Sex
Age
Height
Weight
Team
NOC
Games
Year
Season
City
Sport
Event
Medal

NOC
Region
Notes

Legend

Some New Data Needed

All New Data Needed

Seperated User data for space

The Games

Games_ID
Year
Season
Event_ID
Country_ID

The Events

Event_ID
Event_name
Sport
Team_Event
Ind_Event
Winter_Sport
Summer_Sport
Year_Start
Year_End

The athletes

ID
Sex
Season

The Games

Games_ID
Games_name
City
Latitude
Longitude
NOC
Language

The Countries

Country_ID
Team_name
NOC
region
GDPR
Population
Latitude
Longitude
Avg_Temp
Language
Year?

The athletes split up

Athlete Women Summer

ID
Name
Age
Height
Weight
Event_ID
Games_ID
Country_ID
Medal

Athlete Men Summer

ID
Name
Age
Height
Weight
Event_ID
Games_ID
Country_ID
Medal

Athlete Women Winter

ID
Name
Age
Height
Weight
Event_ID
Games_ID
Country_ID
Medal

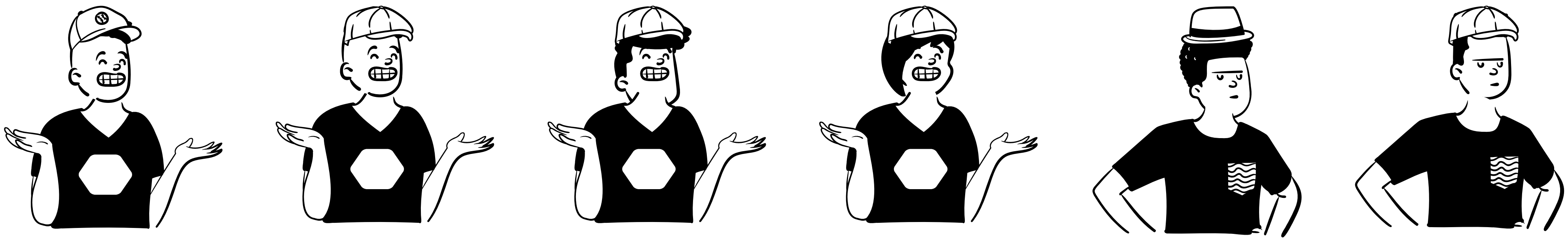
Athlete Men Winter

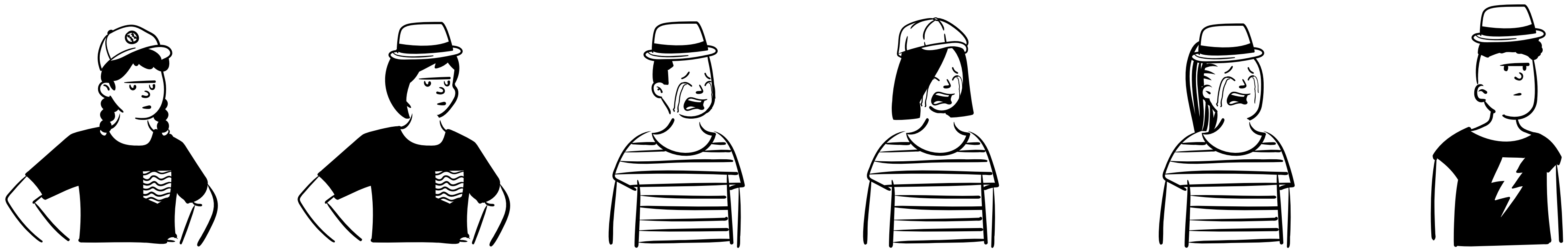
ID
Name
Age
Height
Weight
Event_ID
Games_ID
Country_ID
Medal

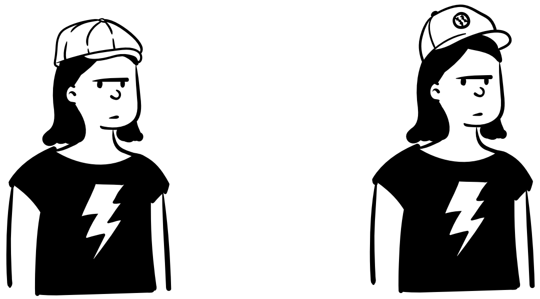
The Data

- More details on how the different tables were created including the code used is presented here:
    - https://thomashsimm.com/sql/pandas/python/olympics/2022/07/29/OlympicsSQL_createDFs.html
    - https://thomashsimm.com/sql/pandas/python/olympics/2022/07/29/OlympicsSQL_createCountryDF.html
    - Creating a country table is a little convoluted as a country can have multiple names and NOC is not a unique identifier (e.g. Russia has 3 NOC values for different time periods for obvious historical reasons)
- Wikipedia was used to get data on population and GDP of different countries. The data was saved as different tabs in the file CountryData.xlsx. For GDP I selected the World Bank Estimate.

- https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)
- https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population







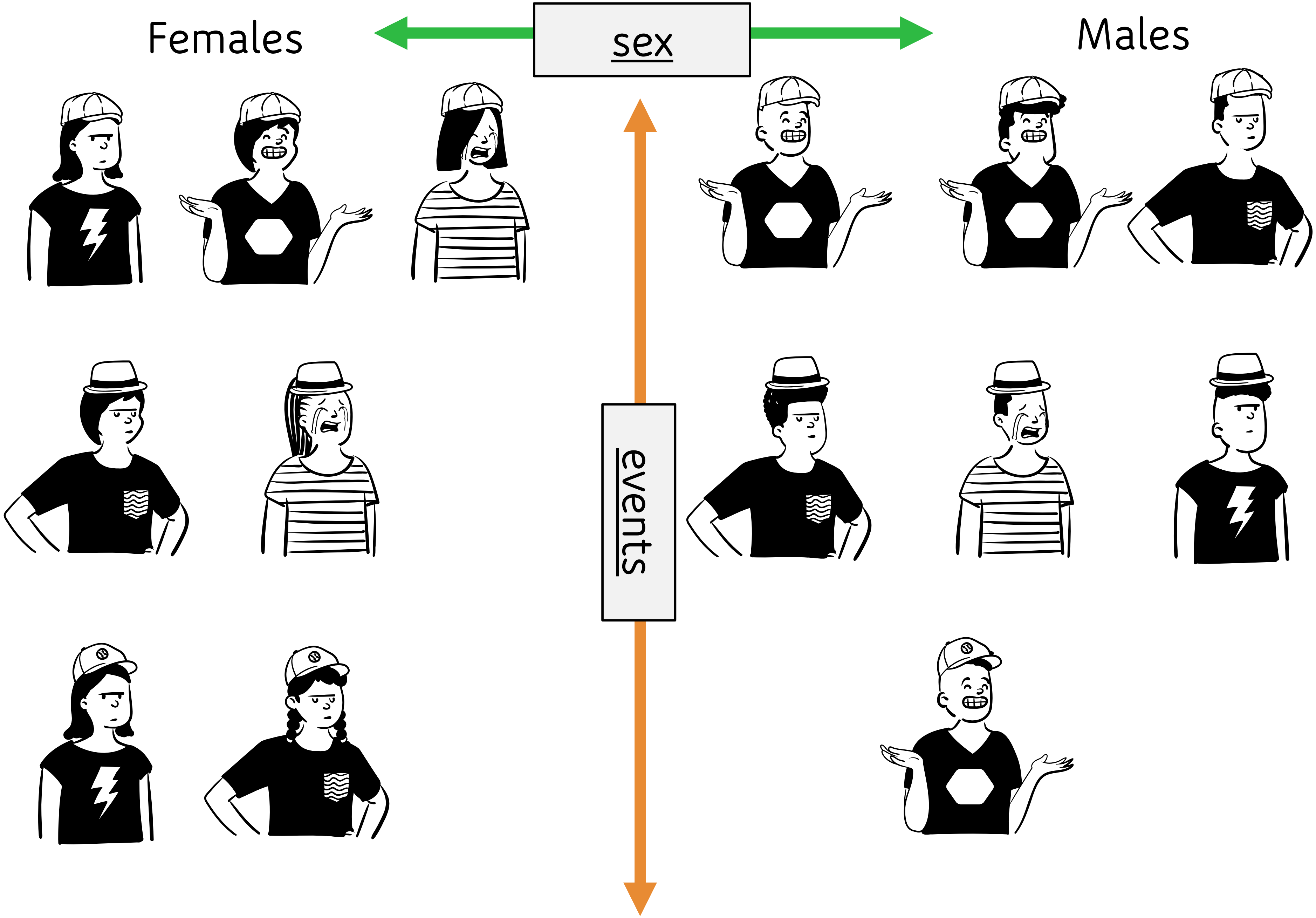
AthleteID	sex	nation	event
1	M	Nation1	Baseball
1	M	Nation1	Flat-cap
2	M	Nation1	Flat-cap
3	F	Nation1	Flat-cap
4	M	Nation2	Panama
5	M	Nation2	Flat-cap
6	F	Nation2	Baseball
7	F	Nation2	Panama
8	M	Nation3	Panama
9	F	Nation3	Flat-cap
10	F	Nation3	Panama
11	M	Nation4	Panama
12	F	Nation4	Flat-cap
12	F	Nation4	Baseball

Visual example of how the data is processed

- Shirt/Facial Expression = nation
- Hat = event
- Hair = sex or M/F

```
SELECT
    event,
    sex,
    COUNT(*)
FROM
    athletes
GROUP BY
    event, sex
```

event	sex	COUNT(*)
Baseball	F	2
Baseball	M	1
Flat-cap	F	3
Flat-cap	M	3
Panama	F	2
Panama	M	3



Females

sex

Males

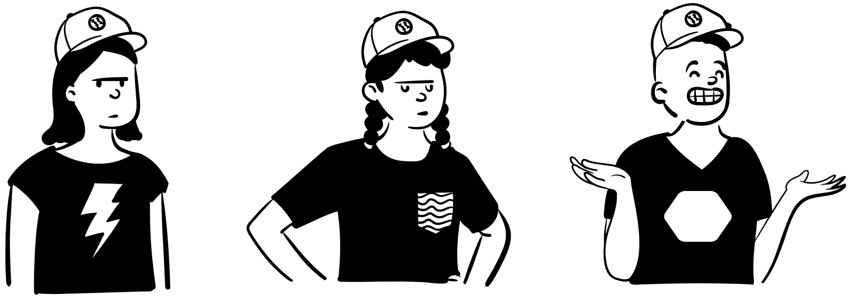
events

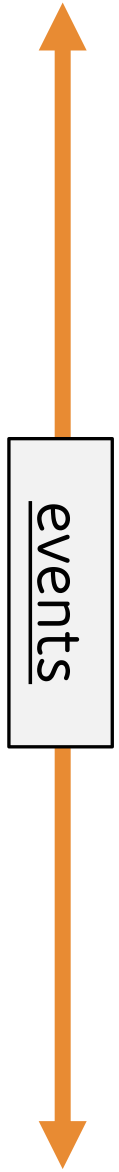
```
SELECT
  event,
  COUNT(*),
  SUM(c)
      FROM
      (SELECT
          event, sex,
          COUNT(*) AS c
      FROM
          athletes
      GROUP BY
          event, sex)
GROUP BY
    event
```

Or for this can do in 1 as

```
SELECT
    event, sex,
    COUNT(*) AS c
FROM
    athletes
GROUP BY
    event
```

event	COUNT(*)	SUM(c)
Baseball	2	3
Flat-cap	2	6
Panama	2	5



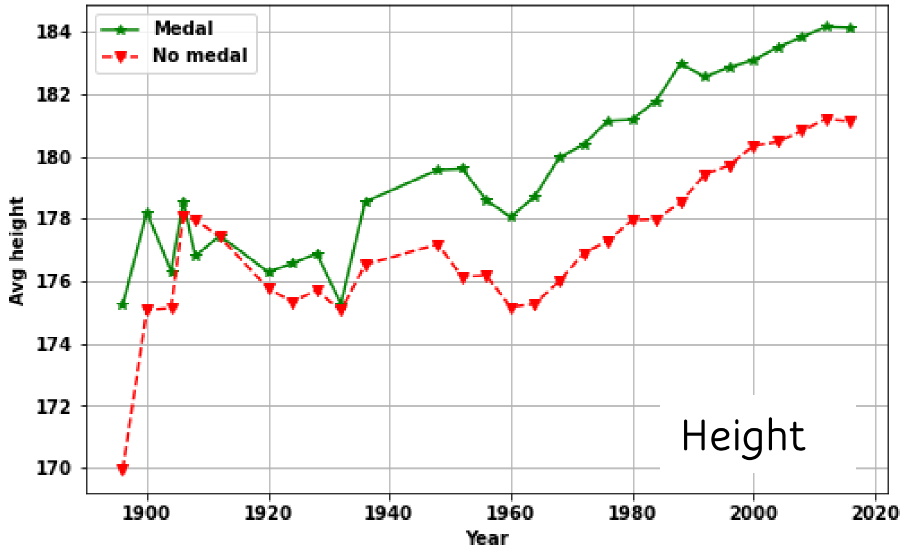


Athletes

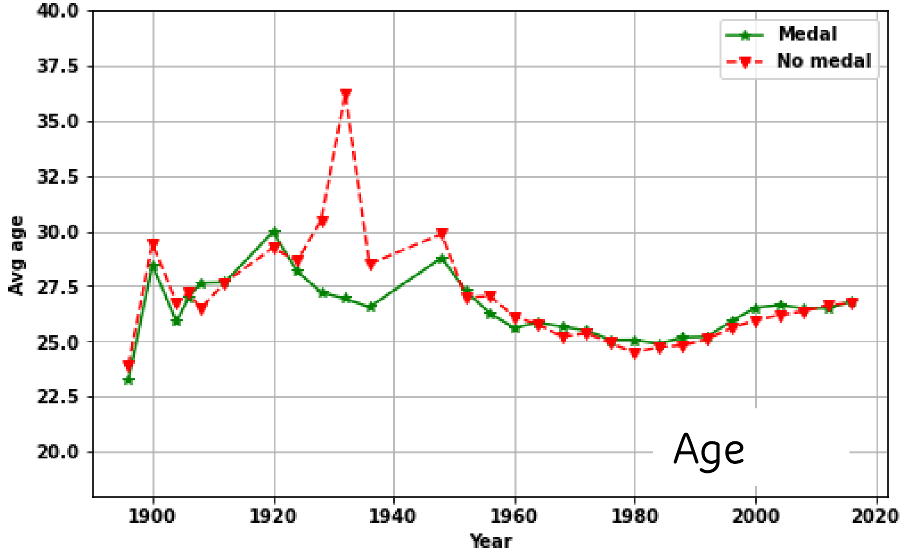
Athlete Stats

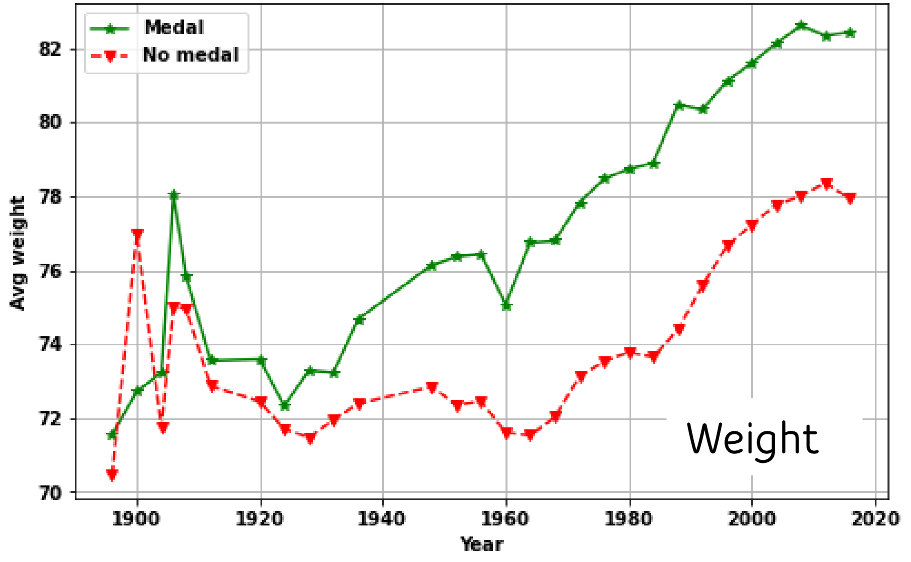
The average height, weight and age of athletes changes over time

- athletes who win medals are heavier than those who don’t
- athletes who win medals are taller than those who don’t
- No obvious effect of age



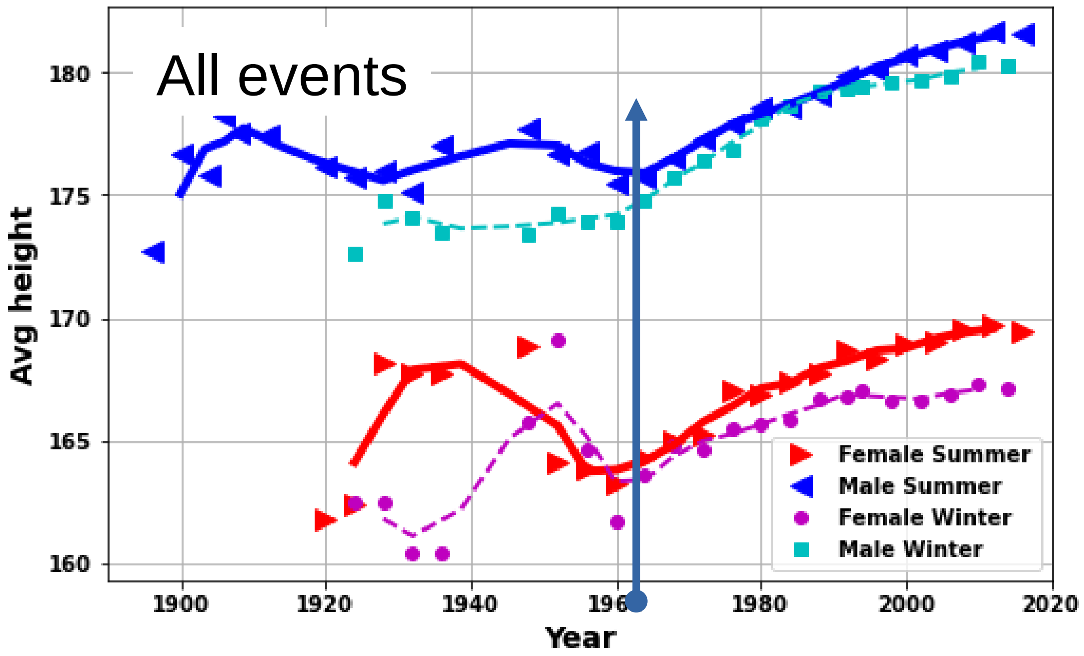
here Male summer athletes are used but the result for female summer athletes show the same trend

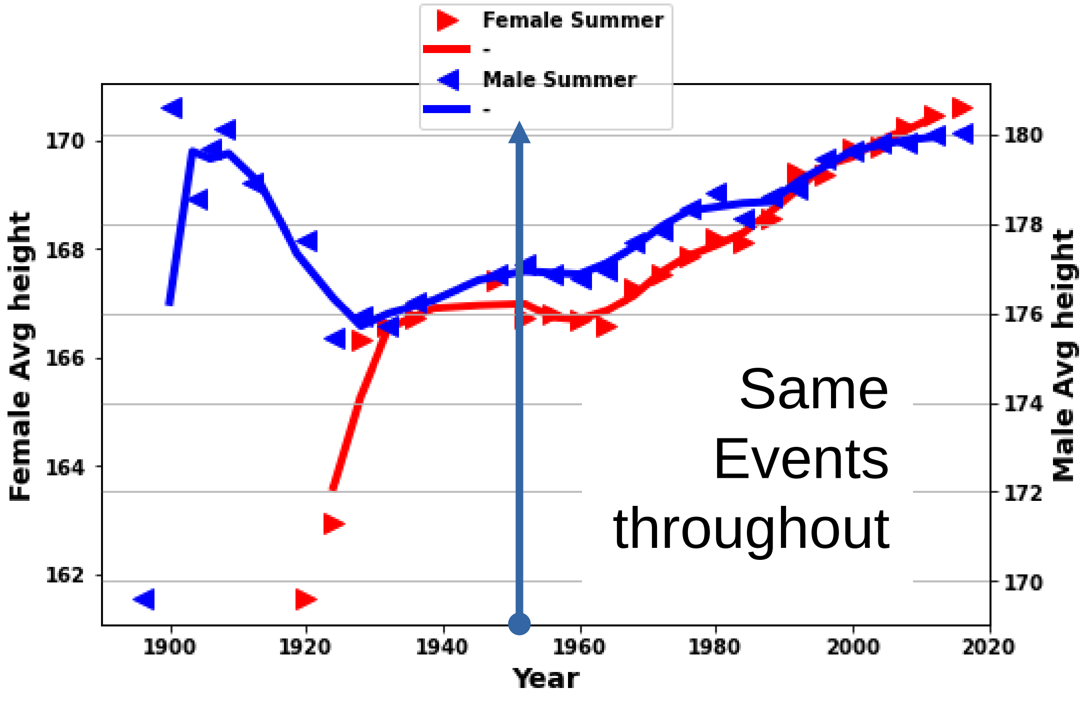




Height

- Increase in height from ~1930
- Acceleration from ~1960
  - The more notable increase from all events may be indicative of what events were introduced
- Stats at lower years less reliable for women
  - N.B. 4 times more female athletes in 1960 compared to 1930



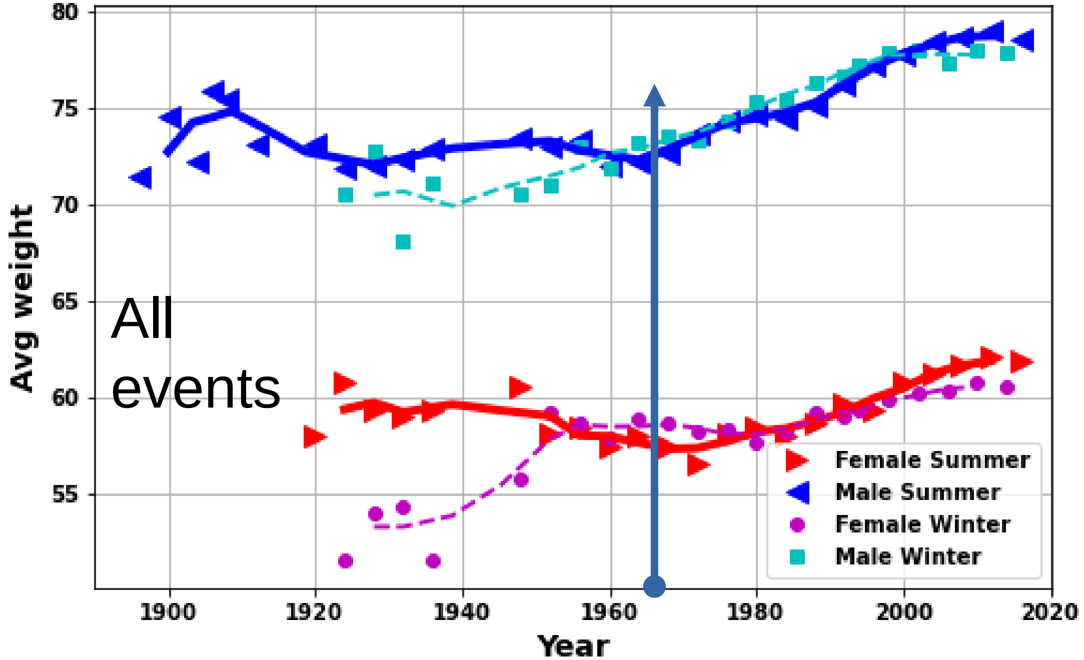


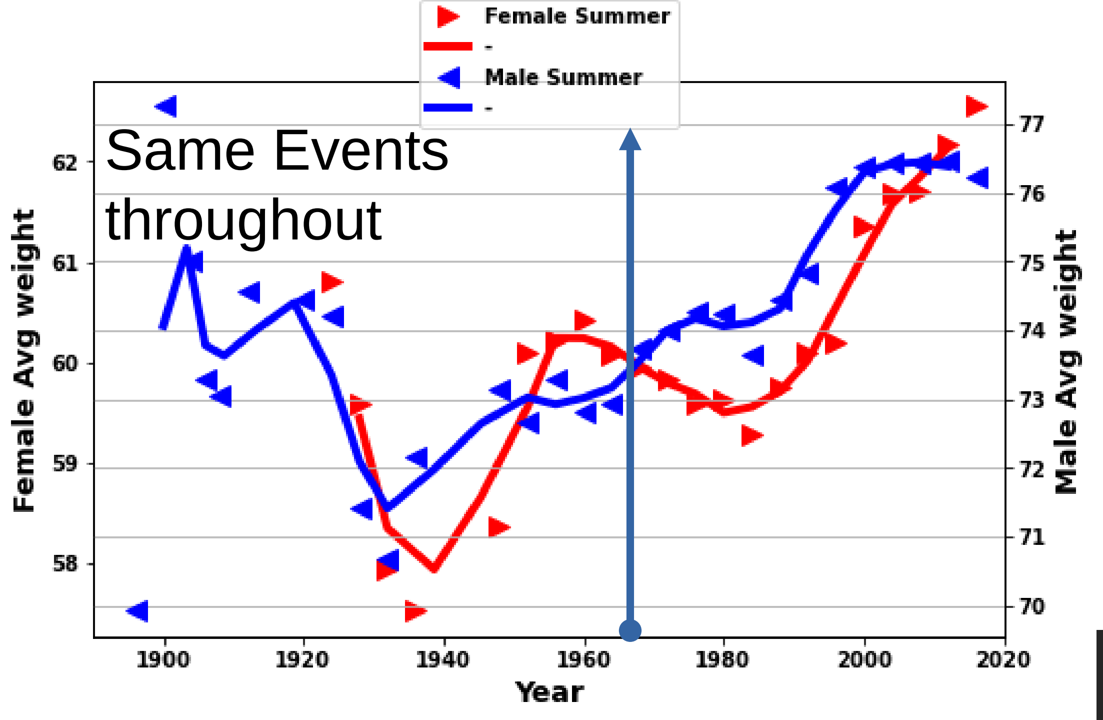
Same Events throughout

This only uses events that occur across a wide range of the Olympics timeframe and averaging is done so each event contribute the same

Weight

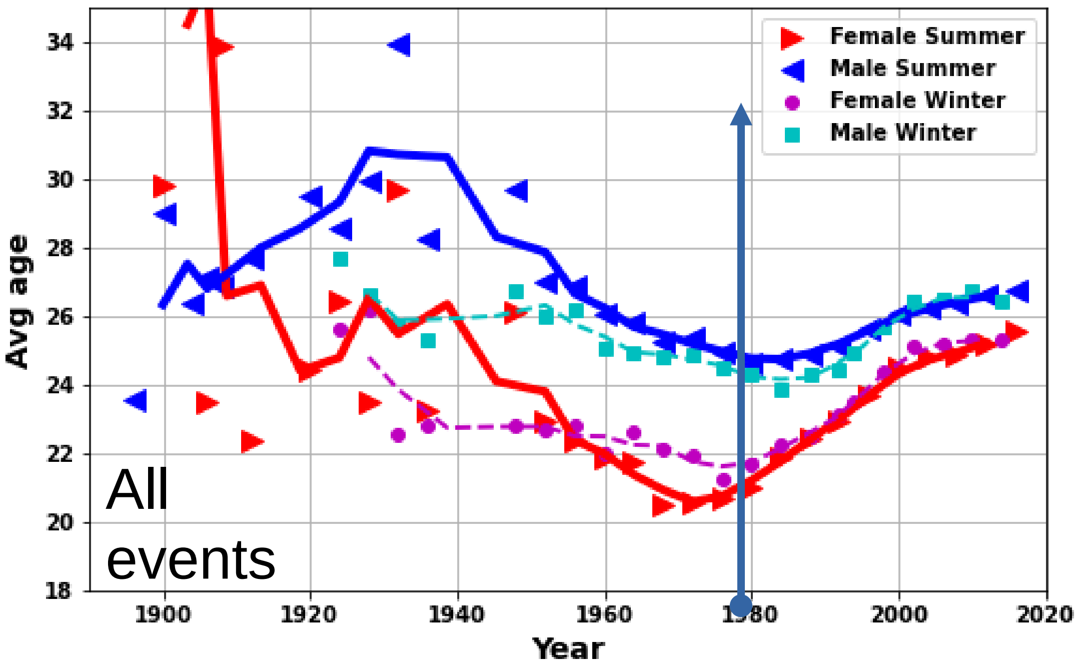
- Increase in weight from ~1930
- The increase from all events may be indicative of what events were introduced after ~1960
    - For example the introduction of sports like water polo has a big effect here
- Unusual local-peak at ~1960 for women

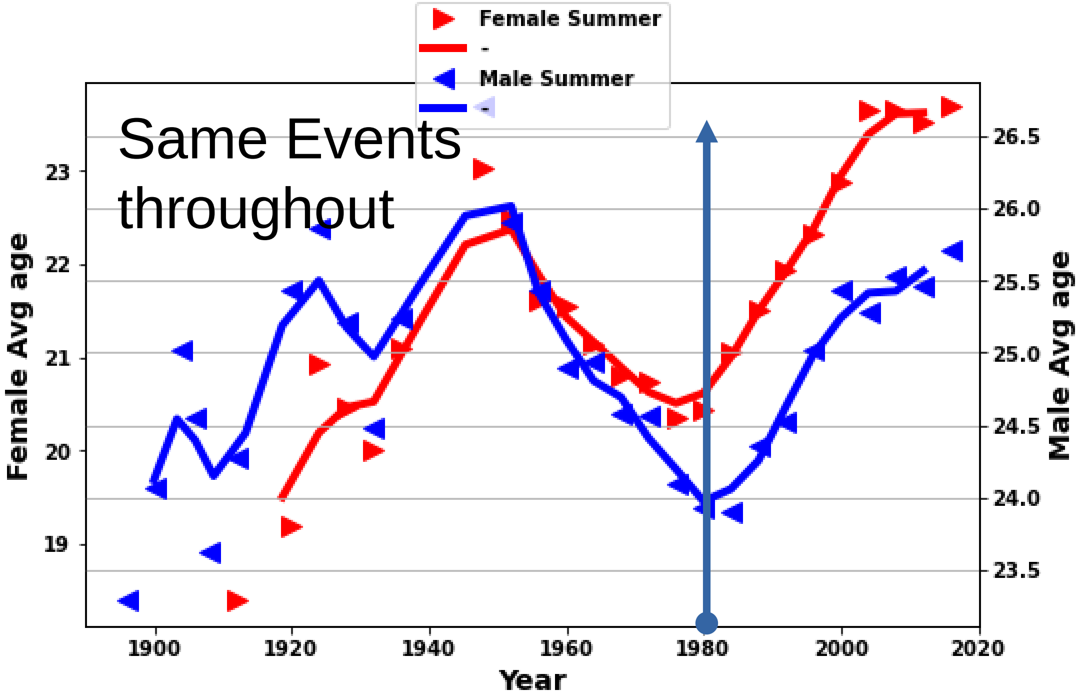

All events


Same Events throughout

Age

- Increase in age of athletes from 1970-1980
    - This would be consistent with a move towards professionalism
    - The amateur rules steadily got relaxed from 1972 and fully abandoned in the 1990
    - But Eastern block were exploiting this from it's inception in the 1950s
- The maximum at ~1950 and the subsequent fall in average age is less obvious
    - Perhaps related to WWII
    - Success in the Olympics becoming more important, alongside athletes using them as a spring board for professional careers (were they can earn money) so the athletes move away from the Olympics when still young

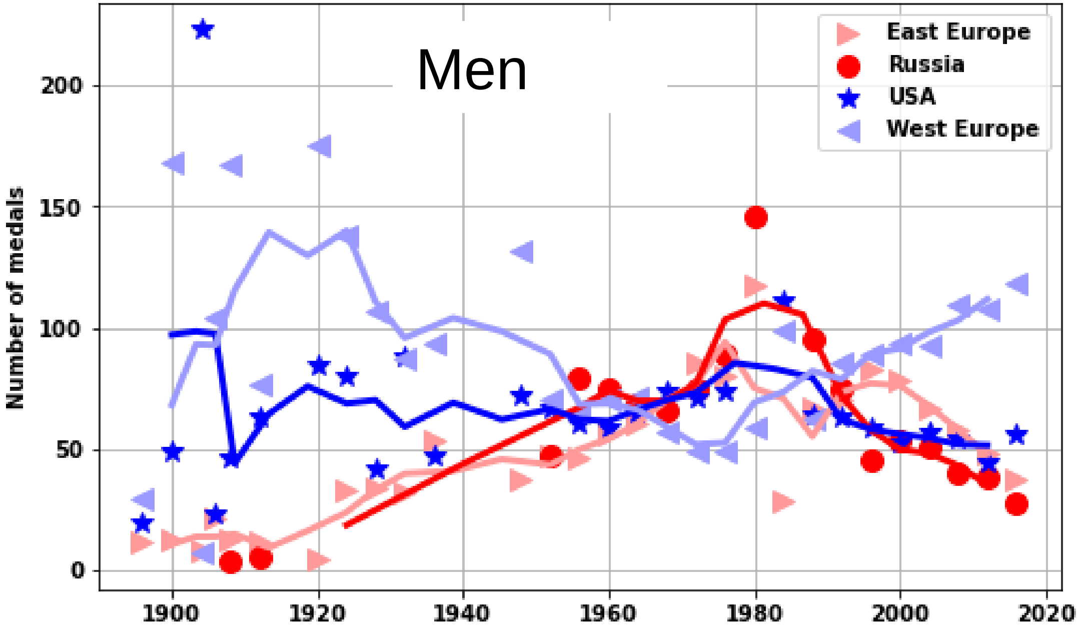




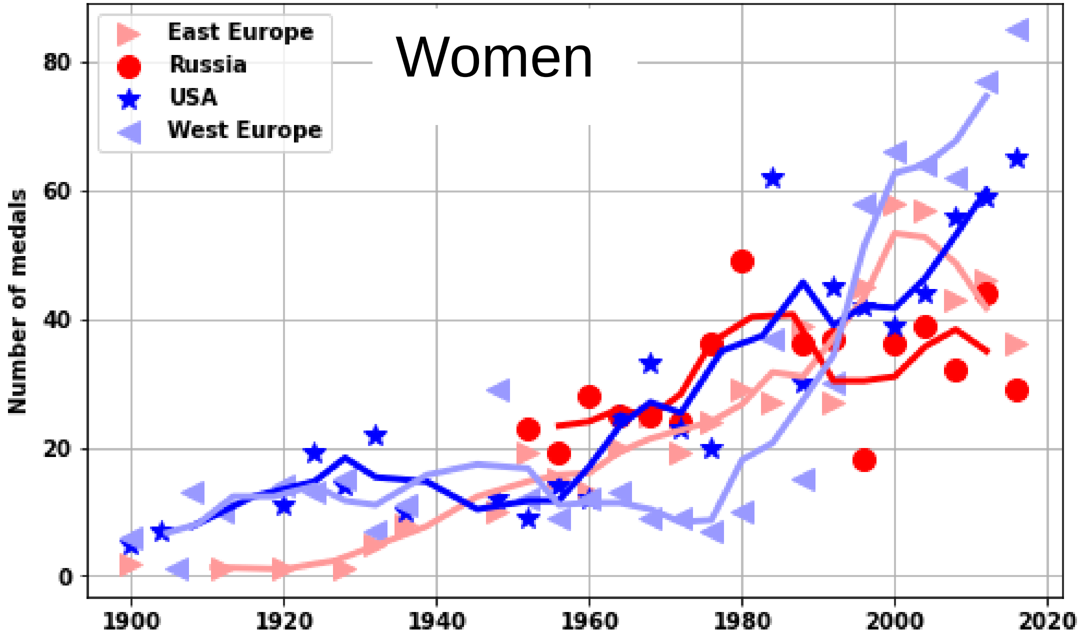
The Cold War

- The USSR (Russia) rejoined the Olympics in 1952
- They dominated the medals table up to USSR’s dissolution in ~1990
- With a similar number of athletes as USA but less than the rest of Europe

▶ East Europe
● Russia
★ USA
◀ West Europe



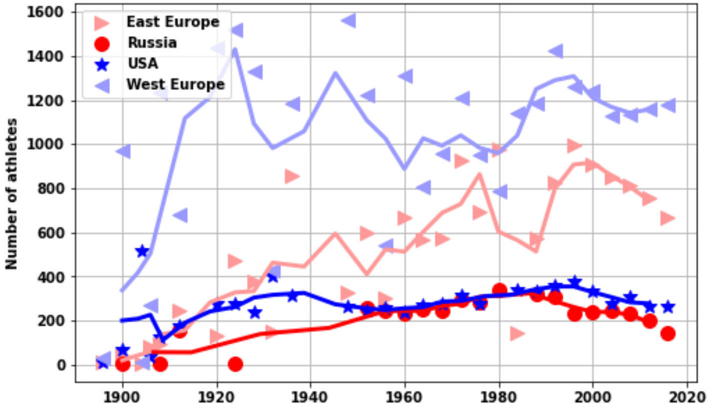
MEDALS

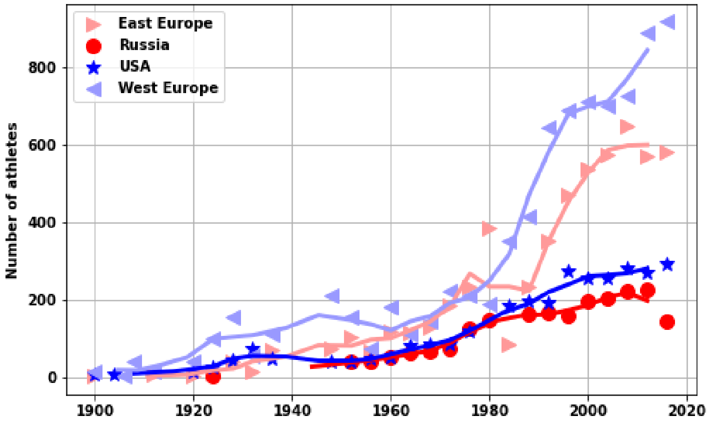


Men

ATHLETES

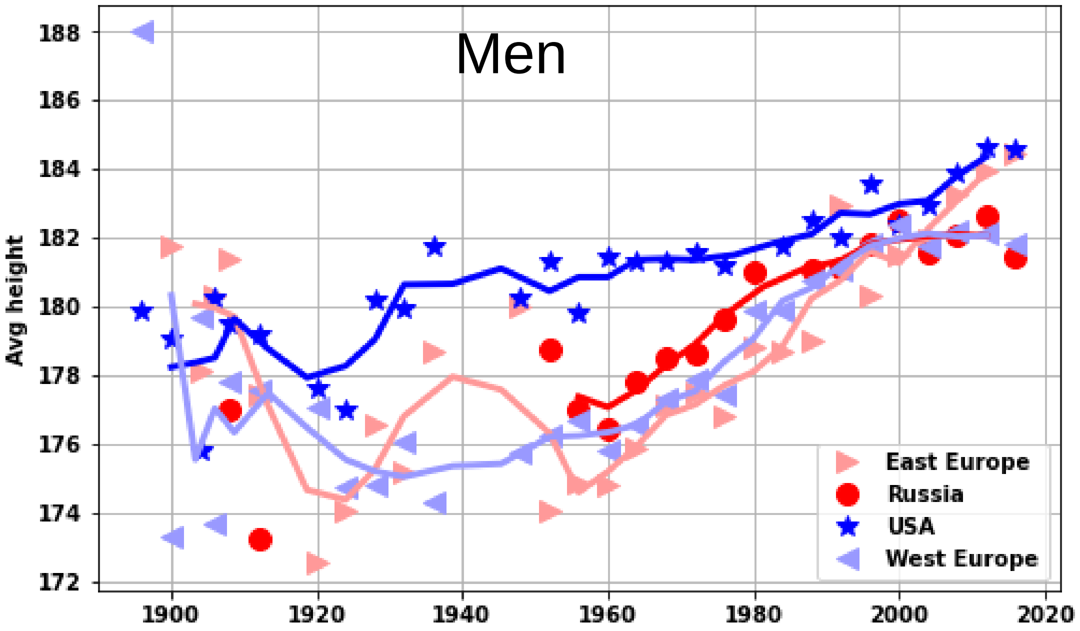
Women

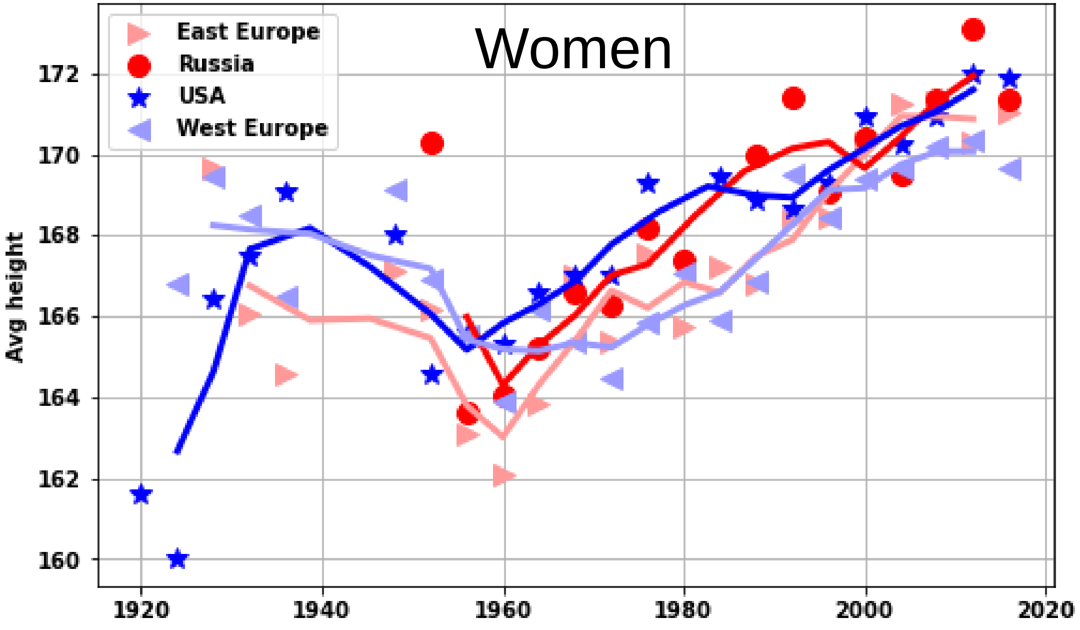




Height

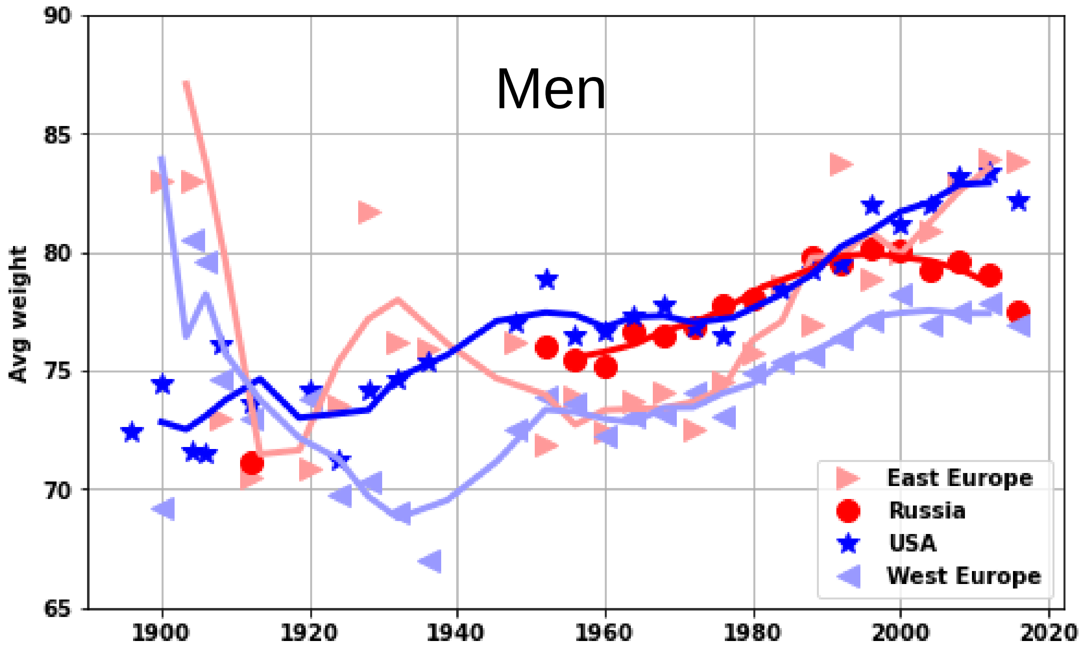
- USA athletes are taller than USSR ones
    - particularly for men
    - but the gap falls and for women they are taller at ~1990
- USA and USSR athletes are taller than the other European nations

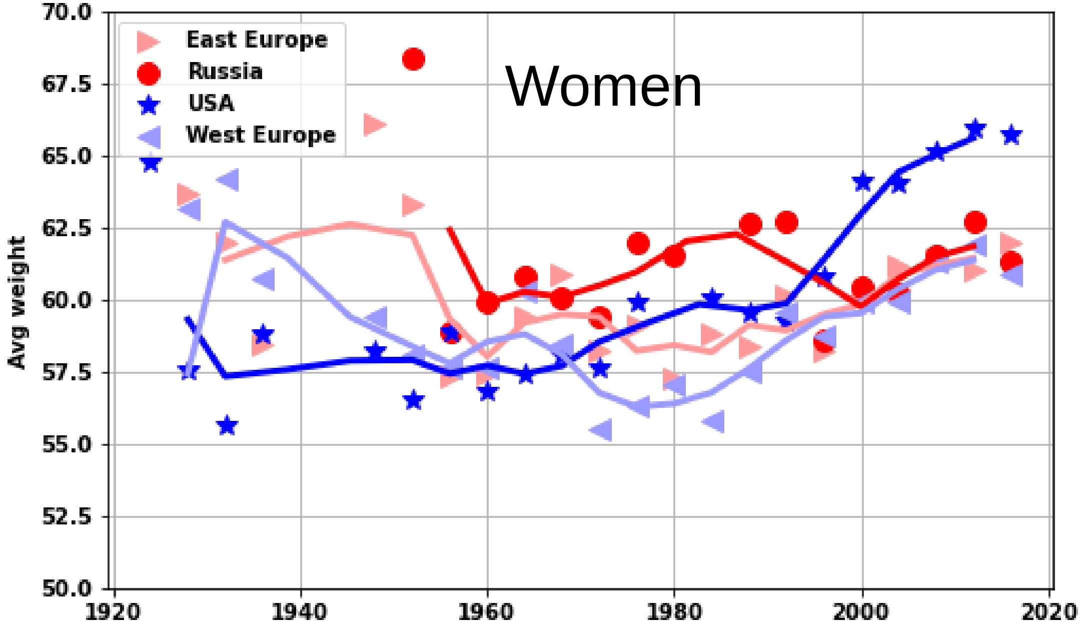




Weight

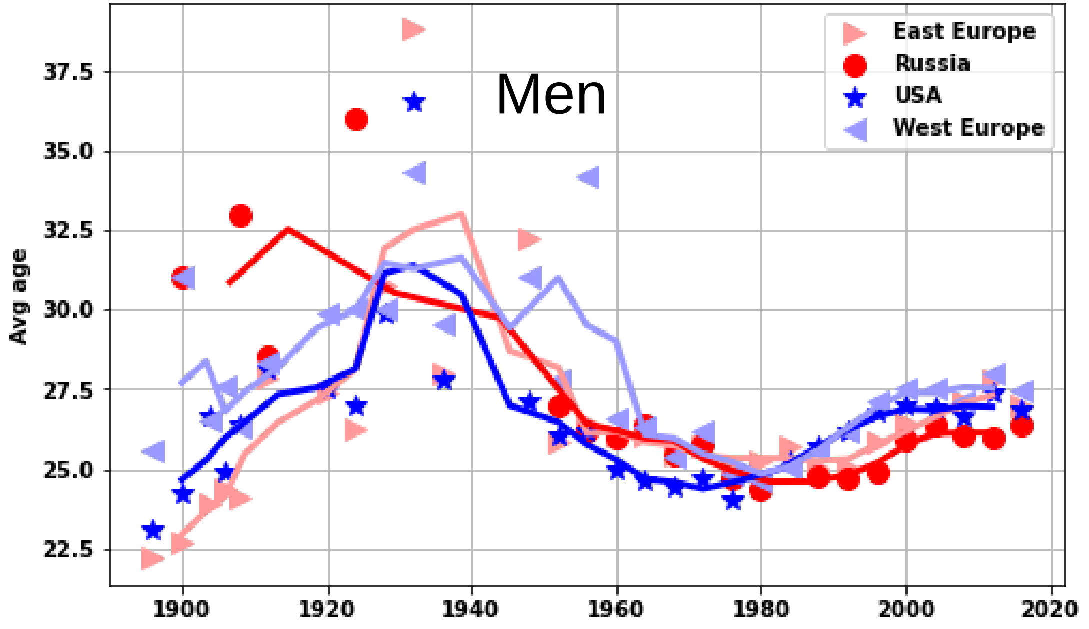
- Russian women are noticeably heavier than USA and other European nations throughout Cold War
  - Then falls below USA after this
- Weight behaviour of men similar to what saw with height
  - USA athletes are heavier but Russian men’s weight overtake USA during the Cold War period

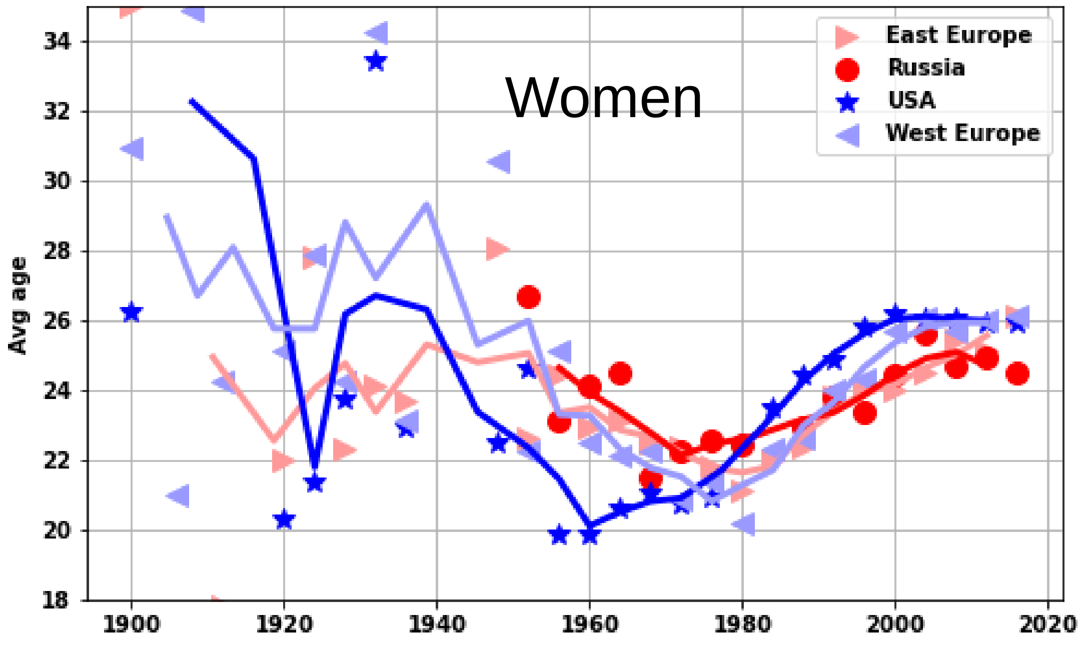




Age

- USA athletes are younger than the other categories prior to 1980
- At ~1980 Russian athletes become younger
- The year at which the age of USA athletes increases (1960-1970) is earlier than for Russian athletes (1980-1990)
- After ~1980 USA are now the oldest average athletes





Nation

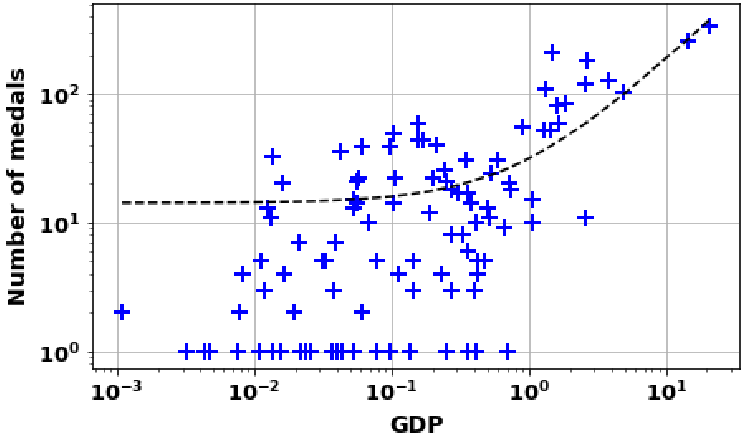
Gross Domestic Profit (GDP)

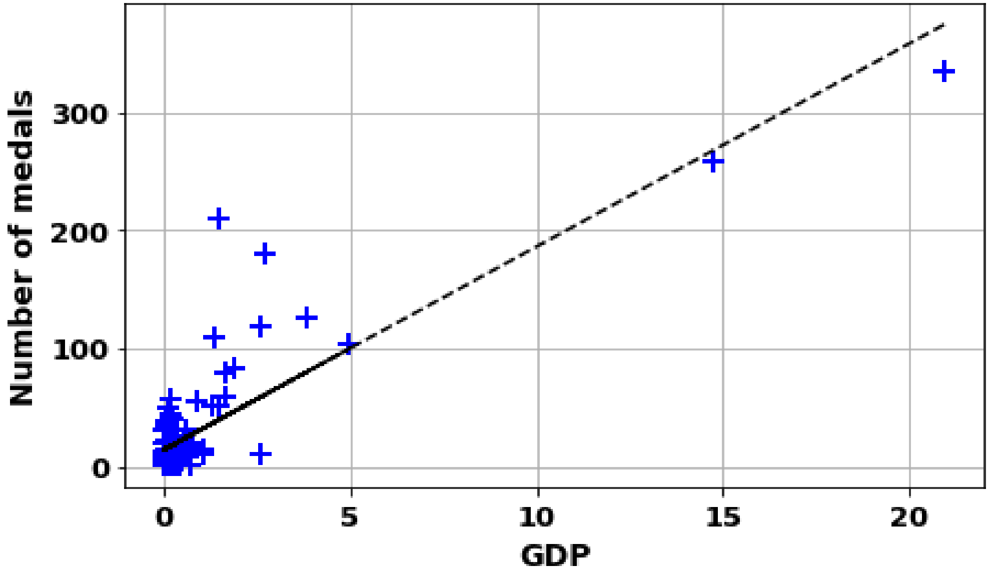
GDP shows a

- Very strong and significant Pearson’s correlation
- Moderate to strong and significant Spearman’s correlation

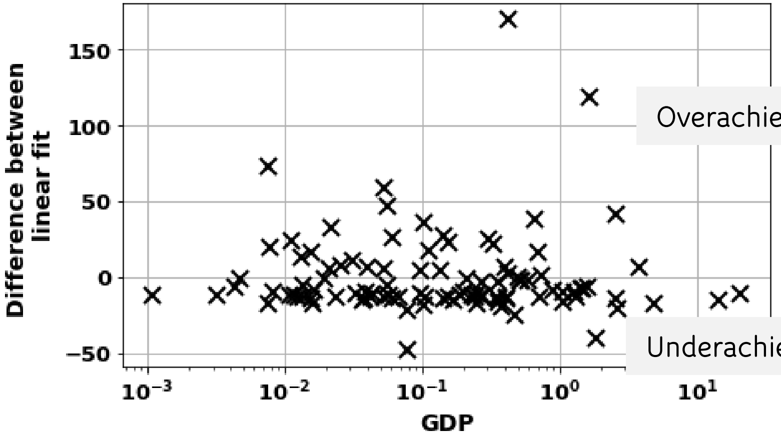
with medals won

Pearson (p-value)	Spearman (p-value)
0.84 (10^{-29})	0.60 (10^{-11})





Underachieve	Overachieve
India	Russia
USA	UK
Saudi Arabia	Australia
Indonesia	France
United Arab Emirates	Germany
Philippines	Ukraine
Israel	South Korea
Mexico	Italy
Austria	Cuba
Chile	Kazakhstan



Overachieve

Underachieve

2008-2016

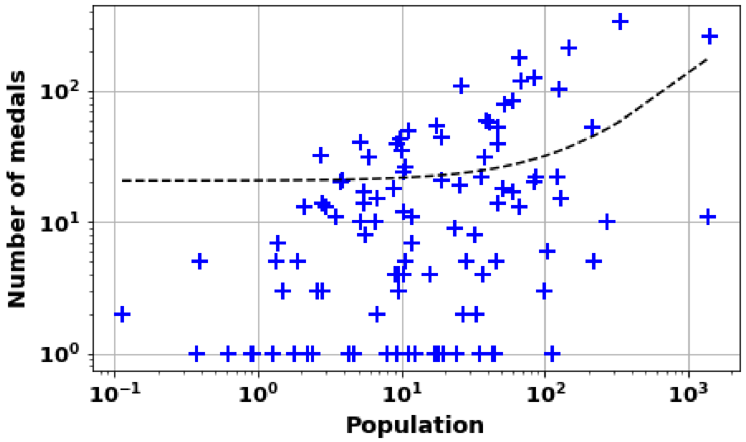
Population

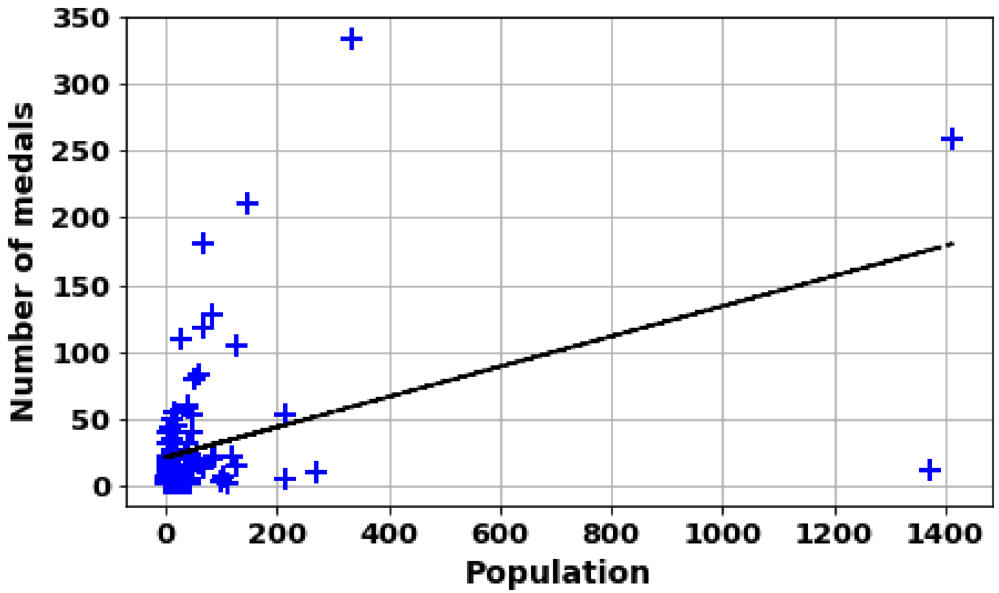
Population shows a

- Moderate and significant Pearson's correlation
- Moderate and significant Spearman's correlation

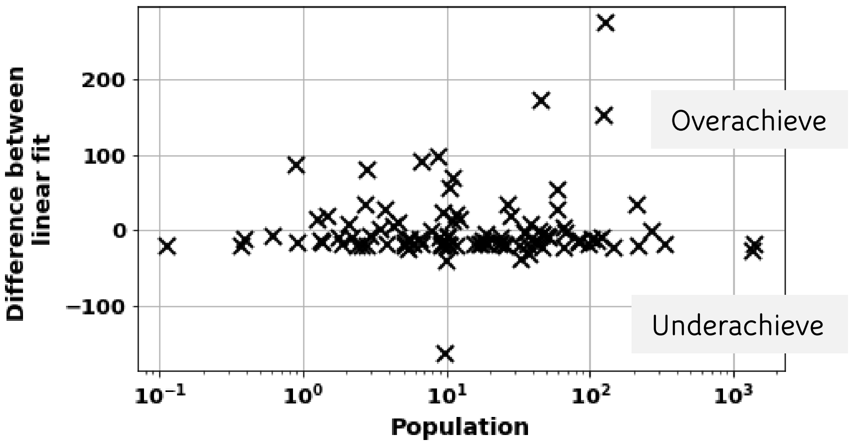
with medals won

Pearson (p-value)	Spearman (p-value)
0.42 (10^{-5})	0.41 (10^{-5})





Underachieve	Overachieve
India	USA
Indonesia	Russia
Nigeria	UK
Philippines	Germany
Vietnam	France
Egypt	Australia
Sudan	China
Uganda	Japan
Saudi Arabia	Italy
Cameroon	South Korea



2008-2016

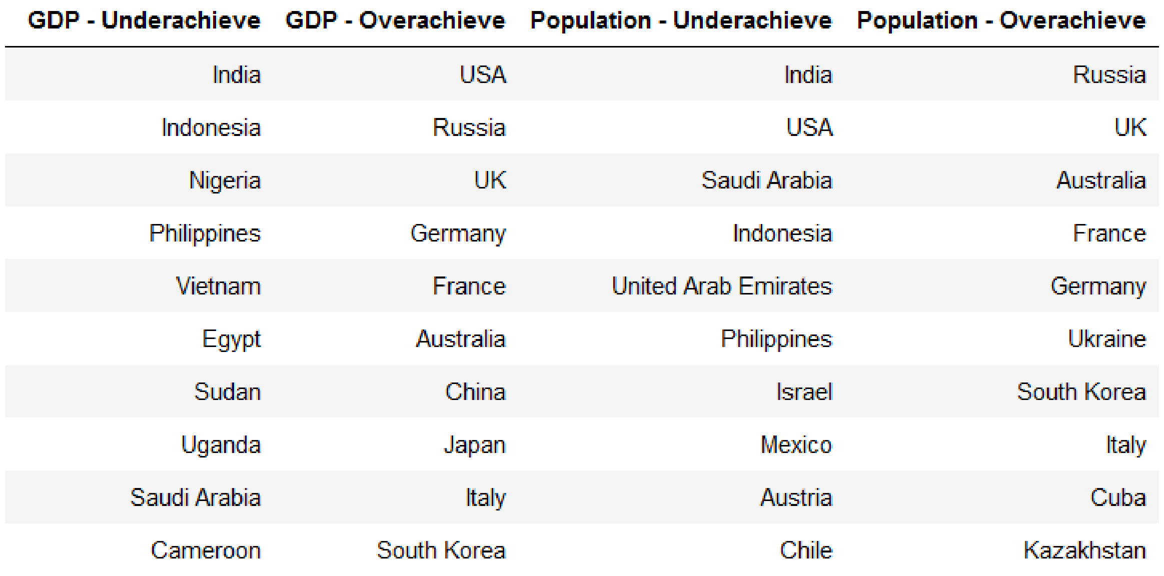
Summary GDP/Population

GDP and Population both show correlation with number of medals a nation obtains
- With GDP showing more correlation

From the countries that **over-achieve** on GDP AND population:

- Similar countries
- Tendency towards European countries
- “Western” or “westernized” countries
  - Australia, Japan

- A split of nations based on their GDP doesn’t increase the correlation
  - Rich/Not rich
- But a split based on whether a country is European or not does increase the correlation



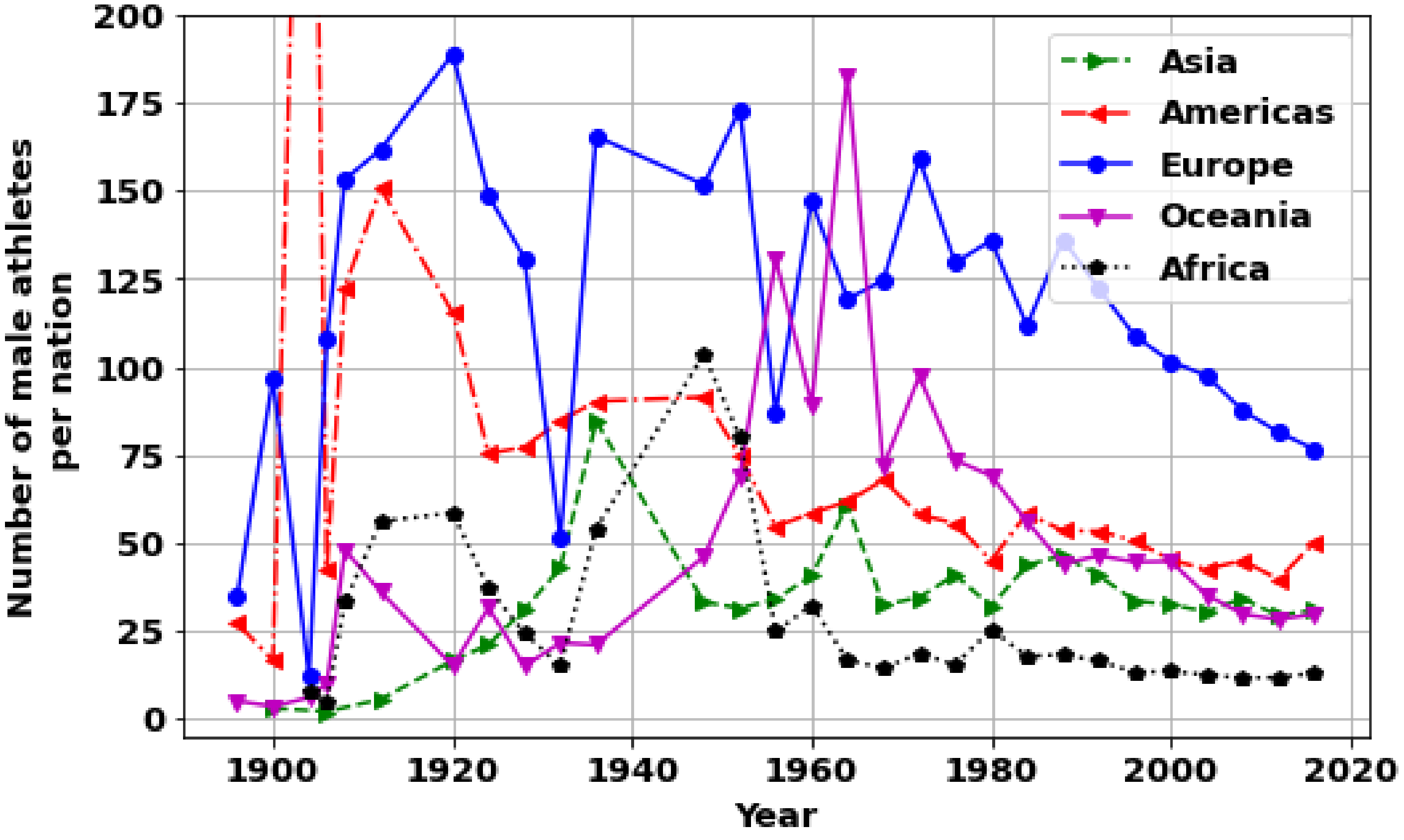
GDP - Underachieve	GDP - Overachieve	Population - Underachieve	Population - Overachieve
India	USA	India	Russia
Indonesia	Russia	USA	UK
Nigeria	UK	Saudi Arabia	Australia
Philippines	Germany	Indonesia	France
Vietnam	France	United Arab Emirates	Germany
Egypt	Australia	Philippines	Ukraine
Sudan	China	Israel	South Korea
Uganda	Japan	Mexico	Italy
Saudi Arabia	Italy	Austria	Cuba
Cameroon	South Korea	Chile	Kazakhstan

		Pearson (p-value)	Spearman (p-value)
	GDP	0.84 (10^{-29})	0.60 (10^{-11})
	Population	0.42 (10^{-5})	0.41 (10^{-5})
Rich/Not rich	GDP	0.84 / 0.32	0.50 / 0.38
Rich/Not rich	Population	0.37 / 0.17	0.30 / 0.15
Europe/Not Europe	GDP	0.80 / 0.94	0.71 / 0.51
Europe/Not Europe	Population	0.93 / 0.49	0.81 / 0.42

Summary GDP/Population

The only useful metric to increase the correlation was to group the countries into those from Europe.

- This is probably due to the similarity of countries within Europe
- And similarities between nations outside Europe (that would have different similarities)
- Countries in Europe have:
    - Similar size and population, and similar GDP
- But also, are similar culturally
    - As shown here European countries have much greater participation at the Olympics (both currently and historically)



Number of athletes in a nations team

Participation at the Olympics is done on merit, and athletes must qualify against athletes from other nations.

So do these 2 correlate:

- the number of athletes per nation attending a particular games
- the number of medals per nation per games

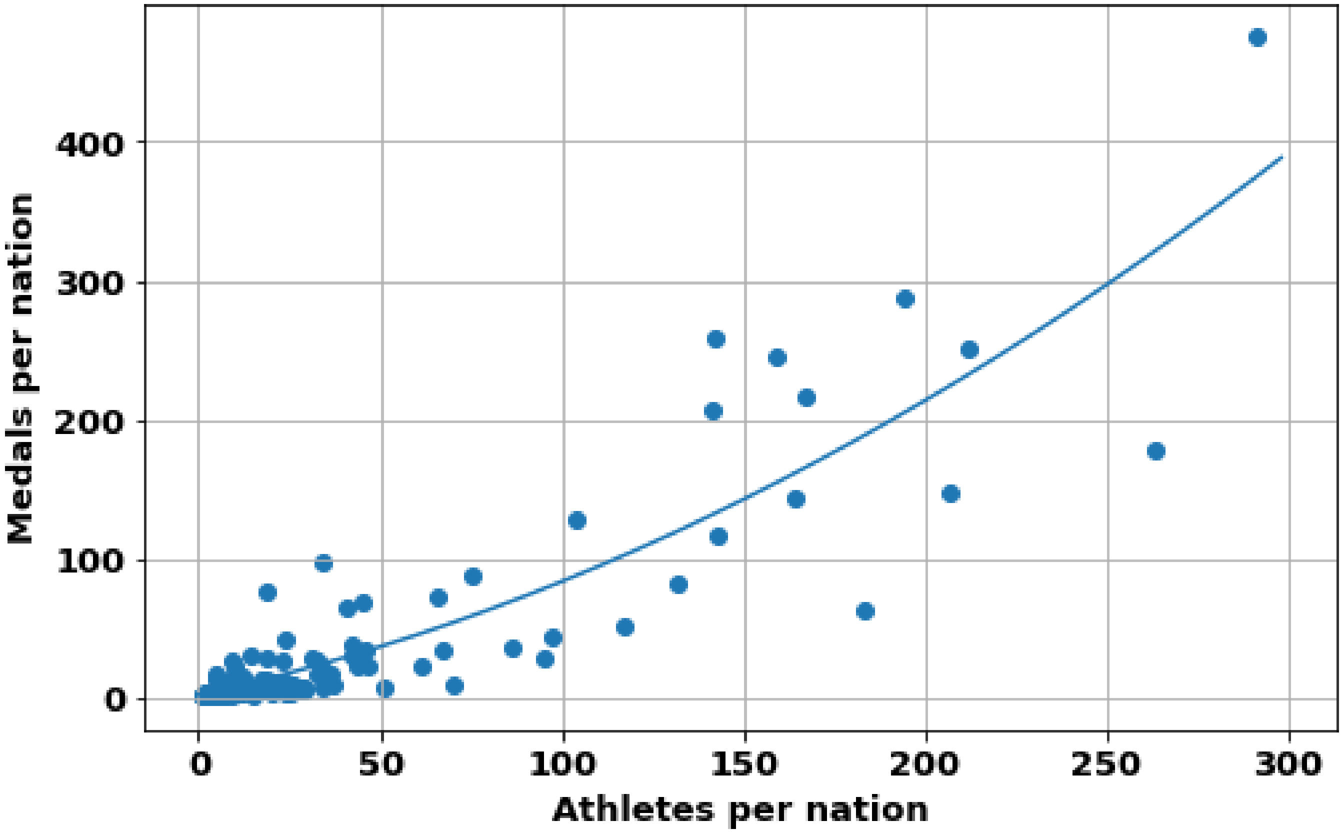
There is a strong correlation between the number of athletes a nation sends and the number of medals they get

Furthermore, the more athletes a nation sends the greater the medals/athlete ratio.

i.e., If a nation sends more athletes, it is more likely that a higher proportion of them will win medals

Spearman correlation = 0.83 (10^{-26})

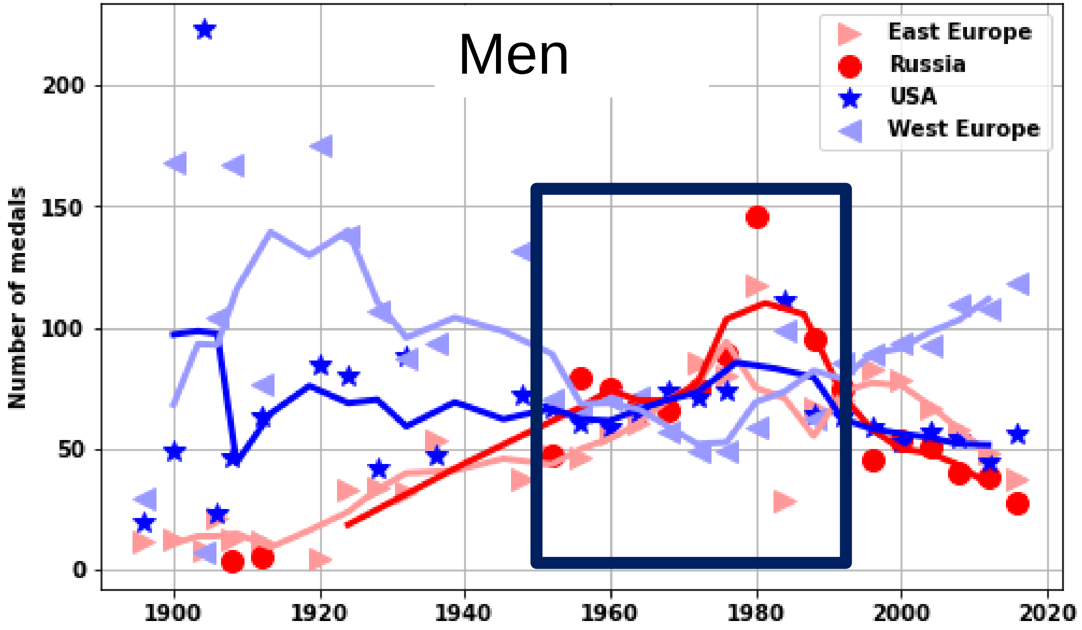
Pearson correlation= 0.87 (10^{-32})



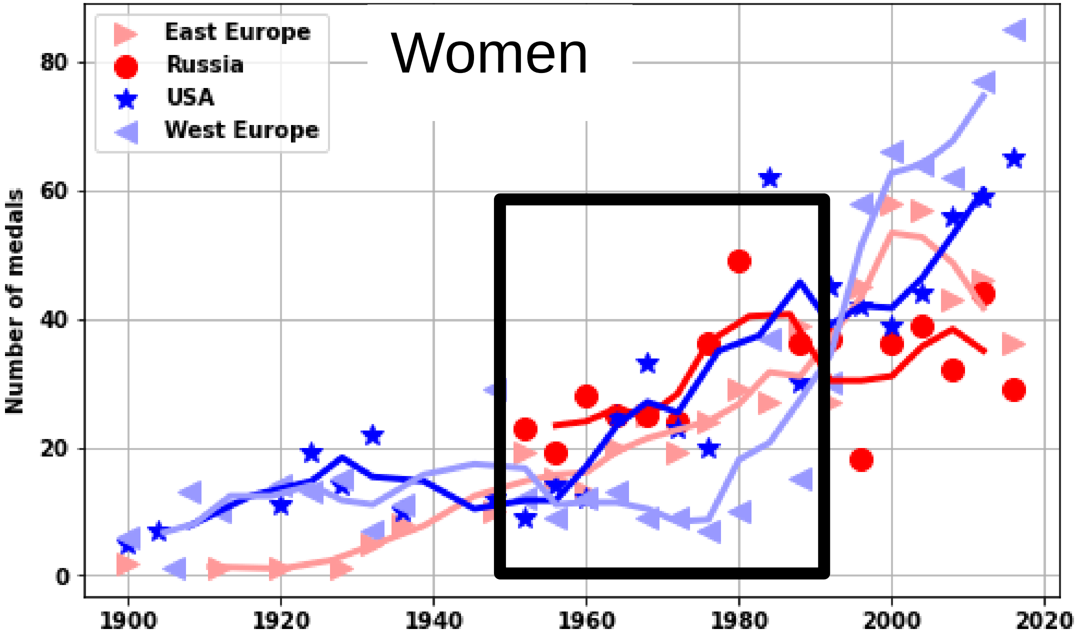
NB I will just use male athletes for simplicity

The Cold War

- The USSR (Russia) re-joined the Olympics in 1952
- They dominated the medals table up to it’s dissolution in ~1990
- With a similar number of athletes as USA but less than the rest of Europe



MEDALS



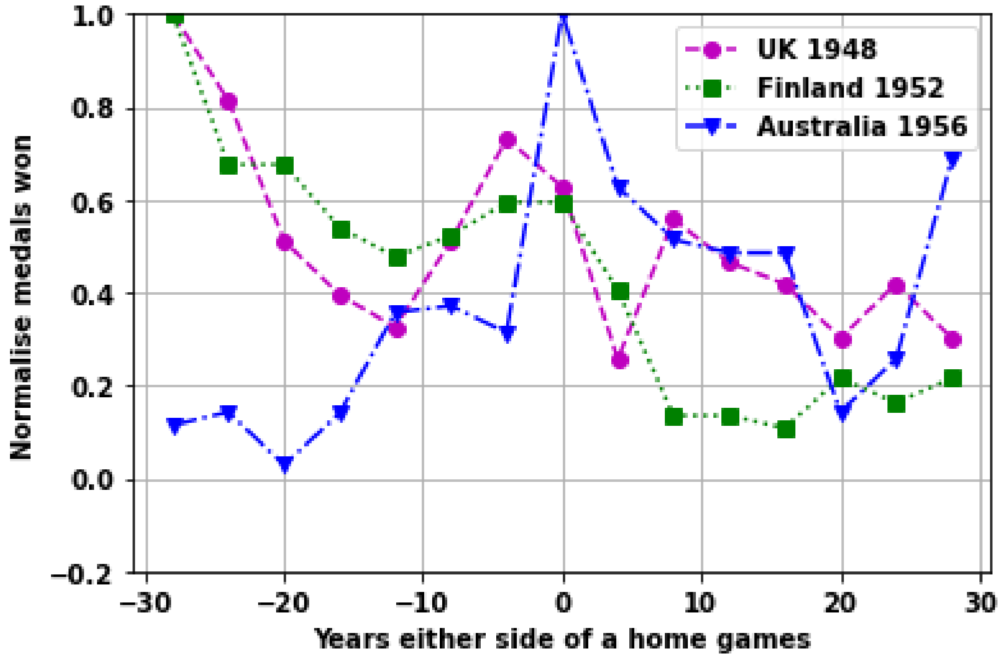
Games

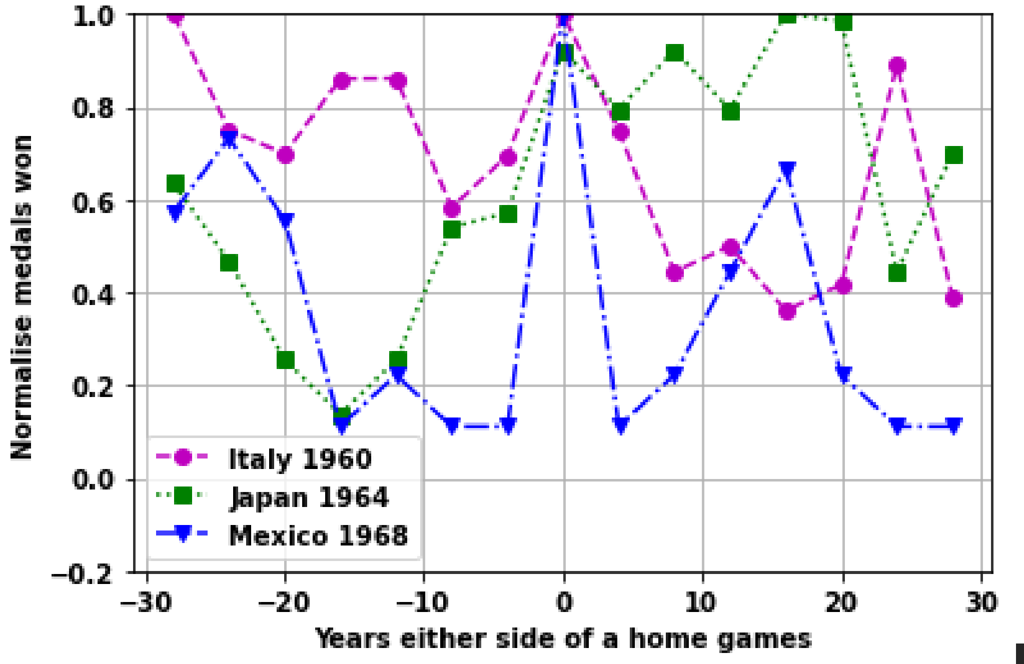
Effect of a home games

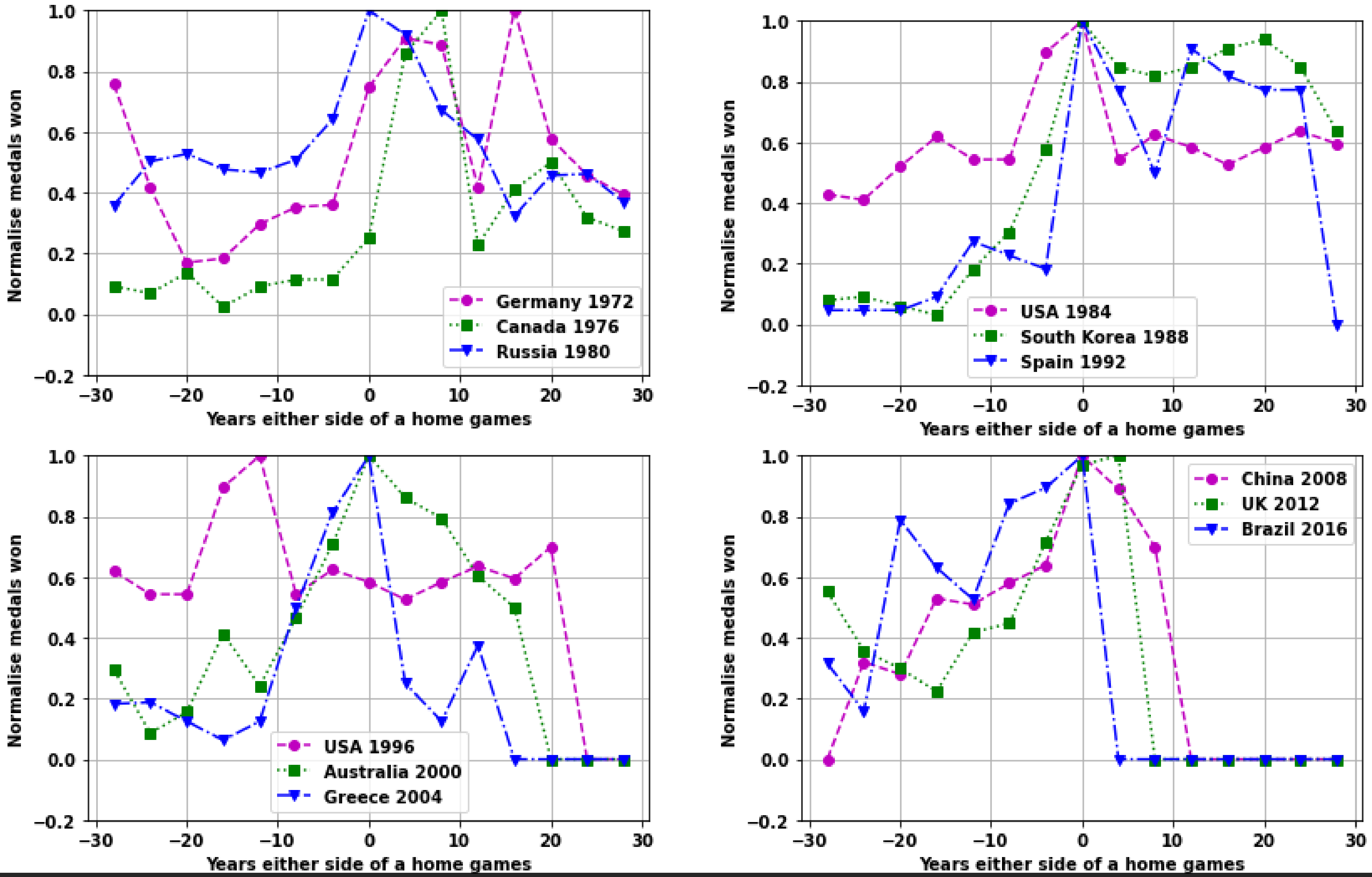
At a home Olympic games a nation should on average obtain more medals than at other games

- Can we quantify this effect?
- Are there any residual effects before and after the games?

- For each games I have interpolated the medals won onto a range of years either side of the home games
    - -32 to +32 years with a step size of 4yrs
- Then added the individual games years/medals data to get an average

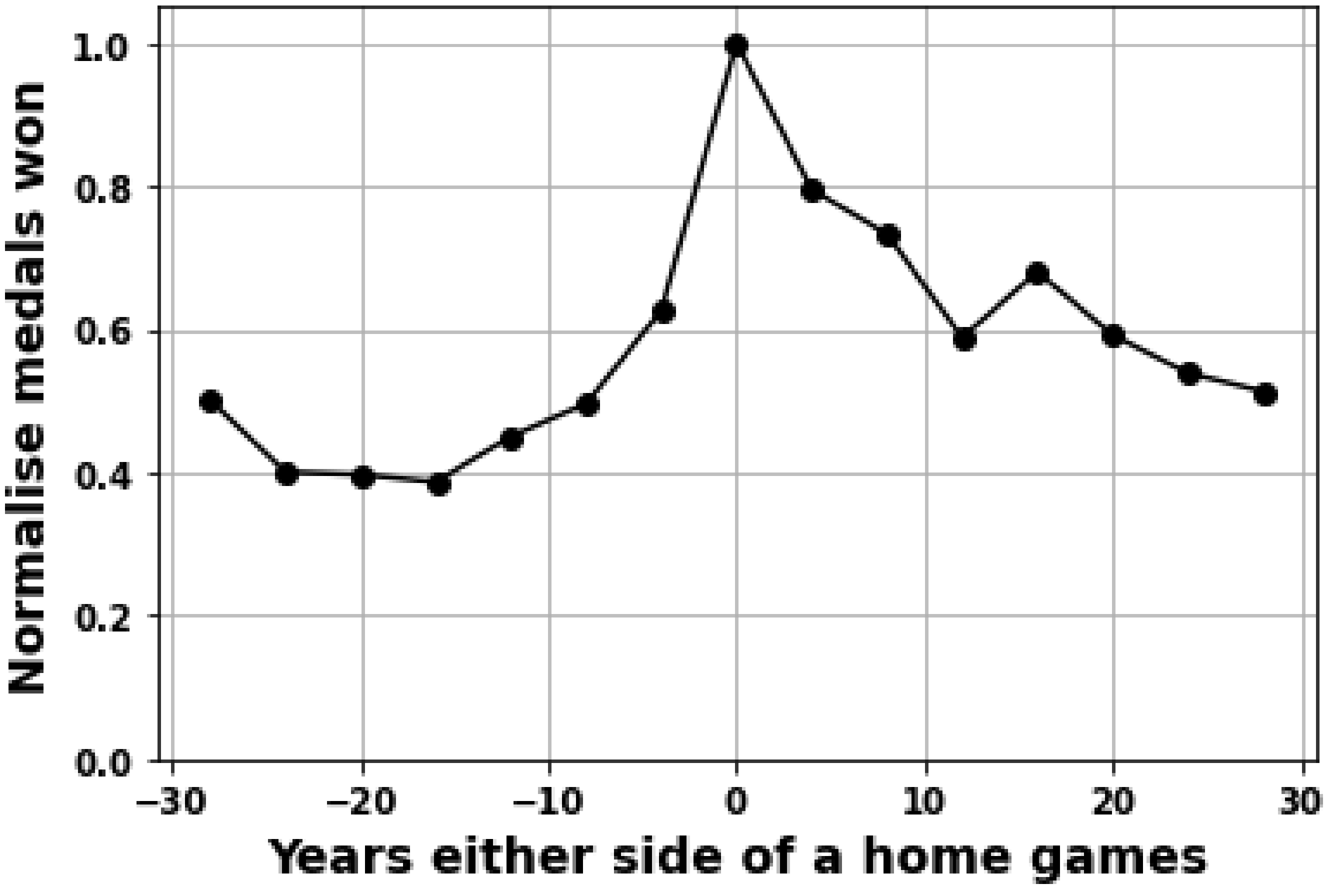





Normalise medals won
Years either side of a home games
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−30
−20
−10
0
10
20
30
Germany 1972
Canada 1976
Russia 1980
USA 1984
South Korea 1988
Spain 1992
USA 1996
Australia 2000
Greece 2004
China 2008
UK 2012
Brazil 2016

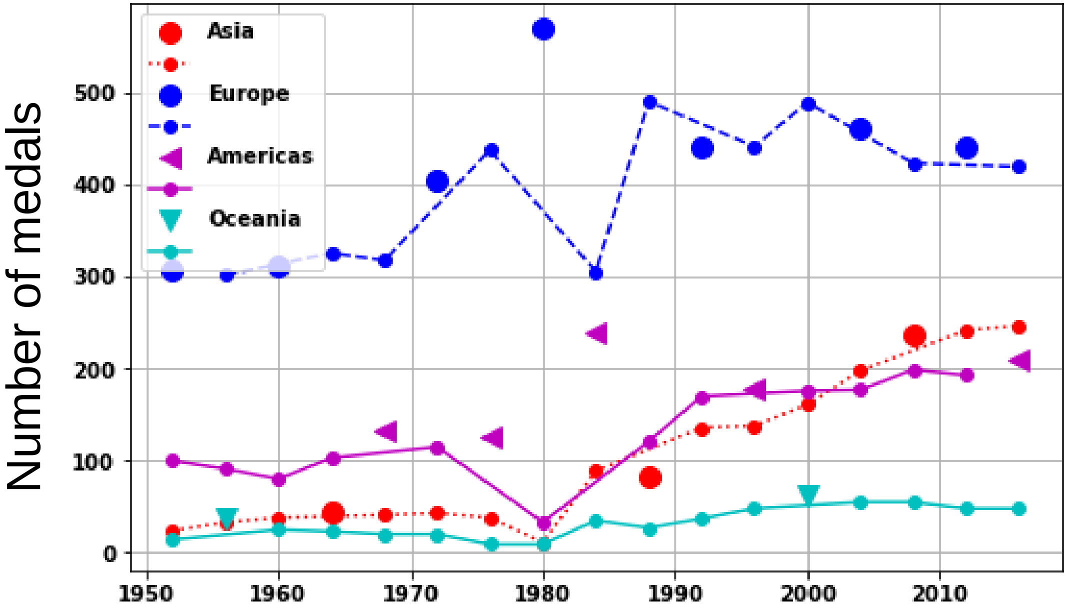
Effect of a home games

- There is an increase in medals obtained for a home games (when a nation competes at a games in their country)
- The home games get ~100% more medals than away from the event
- There is an increase in medals in games surrounding the home games. In the games before the home games the medals obtained are ~60% of the home game and after the game ~80% of the home game. The medals after the home games fall away slower after the games than the increase before the games.



Effect of a home continent games

- The Olympics have mainly been held in Europe and North America. Hence, it would not be feasible to do the same analysis as for home country games
  - Instead, we could look at the difference if a games is a home continent or not
  - But difficult to quantify
    - How to remove effect of the home-Nation uplift
      - Games in Americas and Oceania have a high percentage of athletes that are from the home nations
    - Account for overall changes in medals with time
    - Lack of overall games
  - The best way around the above issues with limited analysis is to use the data from Europe
- From this can estimate the continent effect to be less than 5%.



Continent	Ration medals
Europe	1.06
Asia	1.18
Americas	1.36
Oceania	1.57

Overview

Actions

Recommendations