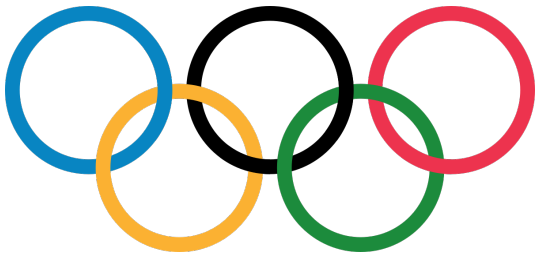
Report: Investigating Kaggle Olympic Dataset

Yasir Din

Data Scientist

March 4, 2019



Contact: yasir@yasirdin.net

Table of Contents

Introduction

Initial Observations

Missing Data

Age

Height & Weight

All-time Performance

Medals per Entrant

Pakistani Medalists

Counting Team Events as 1 Medal

Top 20 Countries Corrected for Teams

Most Successful Olympians

Aside: Medals for Summer & Winter Games

Male & Female Participation

Age Distribution of Athletes

(1) Female & Male Age Distributions Over Time

(2) Female & Male Age Distributions Over Time

(3) Female & Male Age Distributions Over Time

Art Competitions

Berlin 1938

Athlete Height & Weight

Cyclist Weight

Introduction

The work presented here comprises my investigation of the Kaggle dataset: "120 years of Olympic history: athletes and results". Insights and observations are (roughly) presented in the order I discovered them. Starting with macroscopic observations related to the whole dataset; the investigation gradually moves to more detailed insights focusing on a couple of key questions.

Technologies used:

- Python 3.7 (pandas, Matplotlib, seaborn, NumPy)
- Jupyter Notebook

Git Repository:

`https://github.com/yasirdin/kaggle-olympic`

Dataset:

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

Initial Observations

- The Olympics
  - The Summer Olympics started in 1896, and the Winter Olympics in 1924.
  - Until 1992, the Summer and Winter Games were held on the same year. Since then, they have been held on the same 4 year cycle but 2 years apart where: Winter started in 1994 and Summer in 1996.
- `pandas.DataFrame.info` shows missing data in dataset:
  - Height,
  - Weight,
  - Age.
- Interesting `pandas.DataFrame.describe` outputs:
  - Average athlete age = 26 years;
  - Oldest athlete = 97 years!
  - Youngest athlete = 10 years!
  - Average weight = 71 kg;
  - Average height = 171 cm.

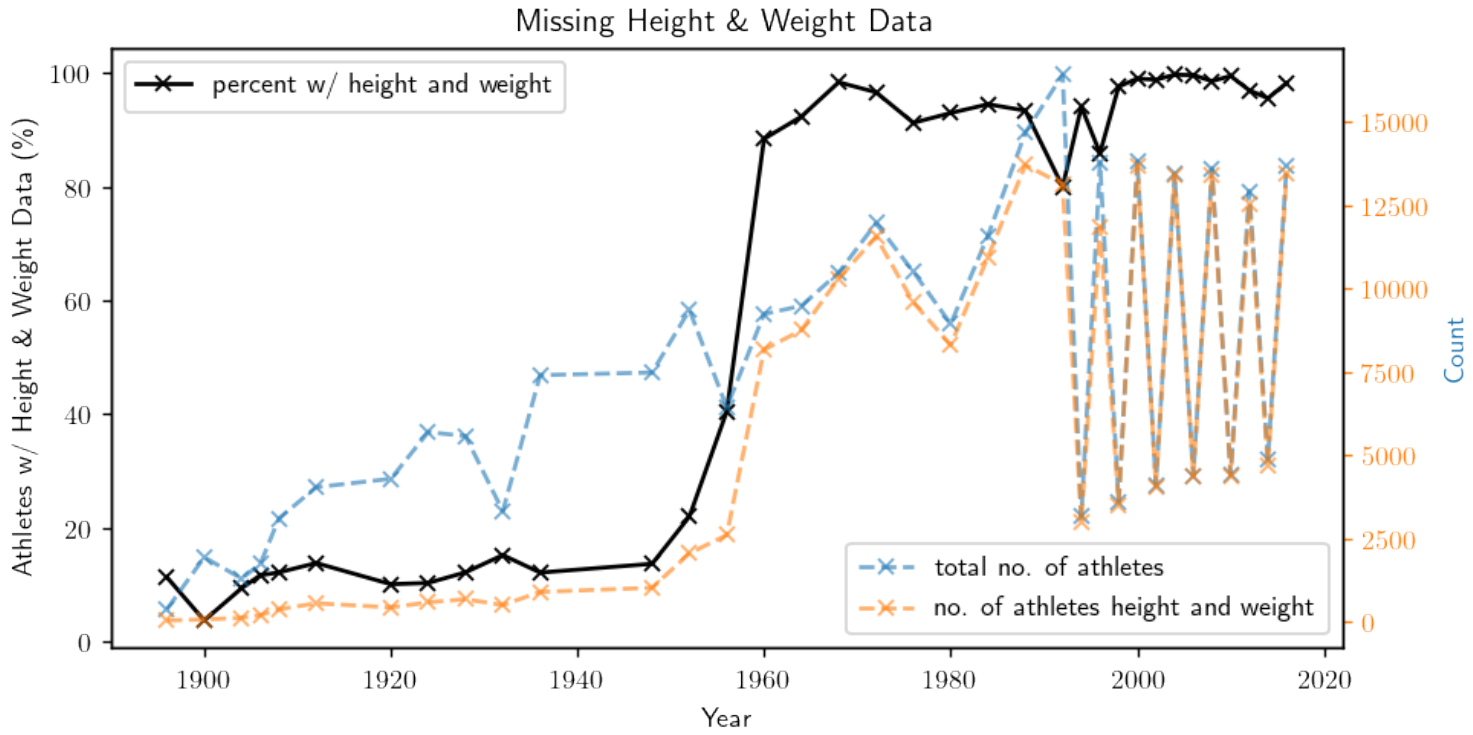
Missing Data (Age)

Left y-axis: percentage of athletes with age data, and right y-axis: raw count of athletes with age data.

Result:

Following 1960, age data for almost all athletes exists (with very few exceptions). Missing data isn't bad overall so drawing rough age statistics from the whole dataset should be fine, however for any detailed analysis *year* $\geq$ 1960 should be taken.

Missing Data (Height & Weight)

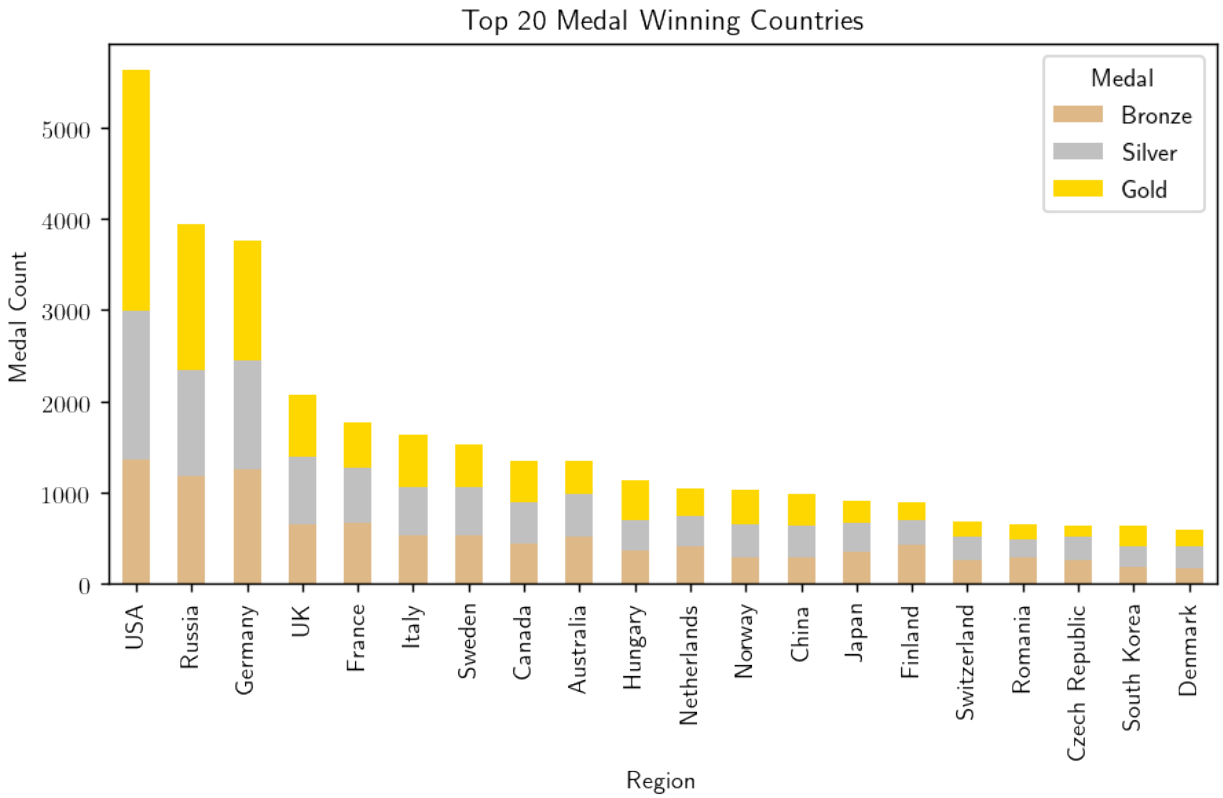


Left y-axis: percentage of athletes with height and weight, and right y-axis: raw count of athletes with height and weight data.

Result:

Following 1960, height and weight data is present for almost all athletes. Not enough data exists before 1960 so any analysis using this data should be done for *year* $\geq$ 1960.

All-time Best Performing Countries

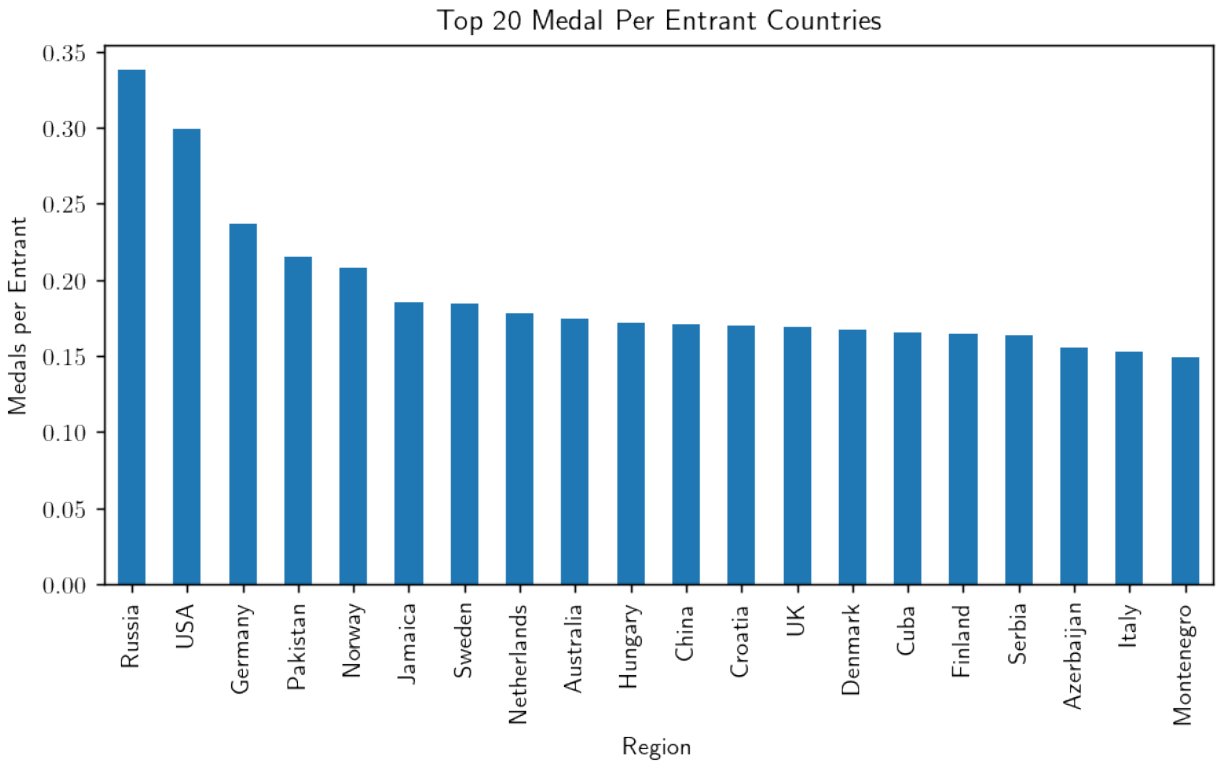


This result was acquired by grouping the dataset by `Region` and not `Country`, so as to include historical changes i.e. Russia == Soviet Union, West / East Germany == Germany, etc.

Thoughts:

USA, Russia and Germany are large countries with high participation in the Olympics. How does this result look when normalised for number of entrants? Let's investigate further...

Medals per Entrant



Top 20 countries with the highest medals per entrant:
$\sum_{medals} / N_{participants}$

Thoughts:

Amongst the other countries, Pakistan is an odd one to see here as they're not known for their Olympic success. Or, perhaps the normalisation has worked and Pakistan have a small number of entrants who have been very successful. Let's find out more...

Pakistani Medalists

Output from checking the events Pakistan have won medals in:

```
Out[53]: Year    Event
         1956.0  Hockey Men's Hockey                          14
         1960.0  Hockey Men's Hockey                          13
                 Wrestling Men's Welterweight, Freestyle       1
         1964.0  Hockey Men's Hockey                          16
         1968.0  Hockey Men's Hockey                          13
         1972.0  Hockey Men's Hockey                          15
         1976.0  Hockey Men's Hockey                          16
         1984.0  Hockey Men's Hockey                          16
         1988.0  Boxing Men's Middleweight                     1
         1992.0  Hockey Men's Hockey                          16
         Name: Event, dtype: int64
```

Thoughts:

Aha! The medals have been counted for whole teams. So in 1956, all 14 players in the hockey team received a medal. This explains the result we saw earlier!

Now, can we find a way to prevent this by counting all team event medal wins as only 1 medal? (Instead of $13 - 16$ in the case of hockey)

Counting Team Events as 1 Medal

Code to achieve this:

In [55]:
```
# taking 1 medal for each event
# NOTE: this doesn't take into account draws for medals
medals_wo_team = merged_df[merged_df['Medal'].notnull()]
                    .groupby(['Year', 'Event', 'Medal'])['region']
                    .first().unstack()
medals_wo_team.head()
```

Out[55]:

Year	Event	Medal	Bronze	Gold	Silver
	Athletics Men's 1,500 metres		France	Australia	USA
	Athletics Men's 100 metres		USA	USA	Germany
1896.0	Athletics Men's 110 metres Hurdles		NaN	USA	UK
	Athletics Men's 400 metres		UK	USA	USA
	Athletics Men's 800 metres		Greece	Australia	Hungary

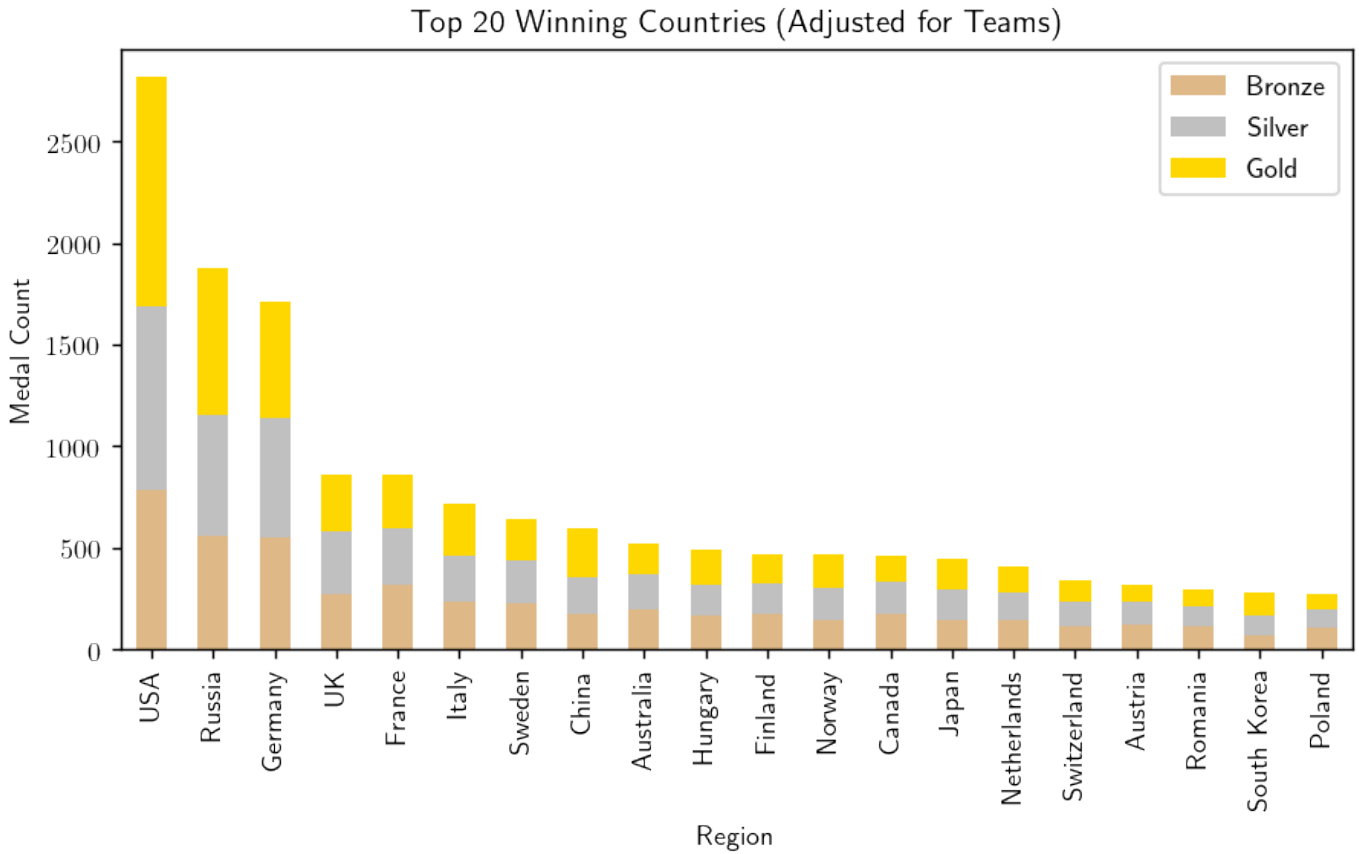
Pakistan Medals after implementing new medal counting system (now correct!):

	Bronze	Silver	Gold	total
Pakistan	4.0	3.0	3.0	10.0

Result:

The new counting system now counts winning teams as 1 medal, as demonstrated with the Pakistan example. After applying this, what do the top 20 countries look like now?

Top 20 Countries Corrected for Teams



Result:

Correcting medal count for team events hasn't had a large effect on the standings. Since all of these team, by virtue of being on top, will participate in roughly the same amount of team events. One exception is China, who has moved up 5 positions (because of relatively more individual events).

Most Successful Olympians

Here are the most successful Olympians of all time:

Name	region	Sport	Bronze	Gold	Silver	total_medals	total_events	medals_per_event
Michael Fred Phelps, II	USA	Swimming	2.0	23.0	3.0	28.0	30.0	0.933333
Larysa Semenivna Latynina (Diriy-)	Russia	Gymnastics	4.0	9.0	5.0	18.0	19.0	0.947368
Nikolay Yefimovich Andrianov	Russia	Gymnastics	3.0	7.0	5.0	15.0	24.0	0.625000
Edoardo Mangiarotti	Italy	Fencing	2.0	6.0	5.0	13.0	14.0	0.928571
Borys Anfiyanovych Shakhlin	Russia	Gymnastics	2.0	7.0	4.0	13.0	24.0	0.541667
Ole Einar Bjrndalen	Norway	Biathlon	1.0	8.0	4.0	13.0	27.0	0.481481
Takashi Ono	Japan	Gymnastics	4.0	5.0	4.0	13.0	33.0	0.393939
Dara Grace Torres (-Hoffman, -Minas)	USA	Swimming	4.0	4.0	4.0	12.0	13.0	0.923077
Jennifer Elisabeth "Jenny" Thompson (-Cumpelik)	USA	Swimming	1.0	8.0	3.0	12.0	17.0	0.705882
Sawao Kato	Japan	Gymnastics	1.0	8.0	3.0	12.0	24.0	0.500000
Aleksey Yuryevich Nemov	Russia	Gymnastics	6.0	4.0	2.0	12.0	21.0	0.571429
Birgit Fischer-Schmidt	Germany	Canoeing	NaN	8.0	4.0	12.0	13.0	0.923077
Ryan Steven Lochte	USA	Swimming	3.0	6.0	3.0	12.0	14.0	0.857143

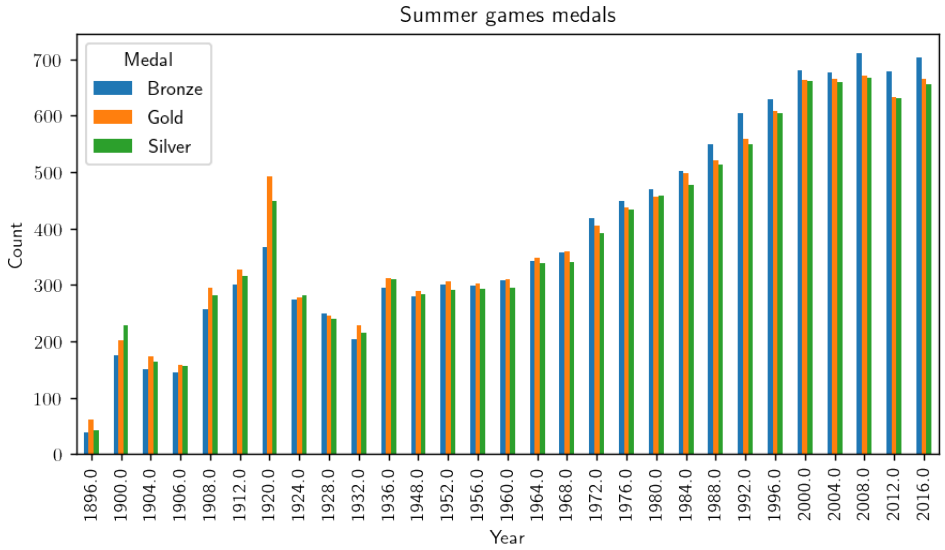
This result was obtained by grouping by `Name`, `Region`, `Sport`, and counting the number of medal wins. Then, `medals_per_events` was added to measure their medal conversion rate.

Thoughts:

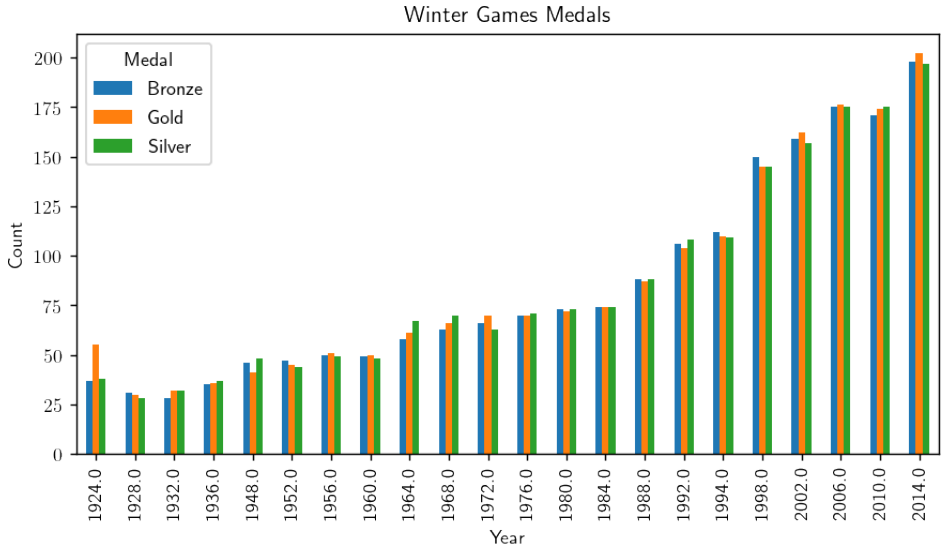
An remarkably high strike rate for Michael Phelps and Larysa Semenivna Latynina; they each win a medal $\sim 94\%$ of the time they compete in an event, with Phelps winning 23 golds out of 28 medals overall!

Aside: Medals for Summer & Winter Games

Summer games:

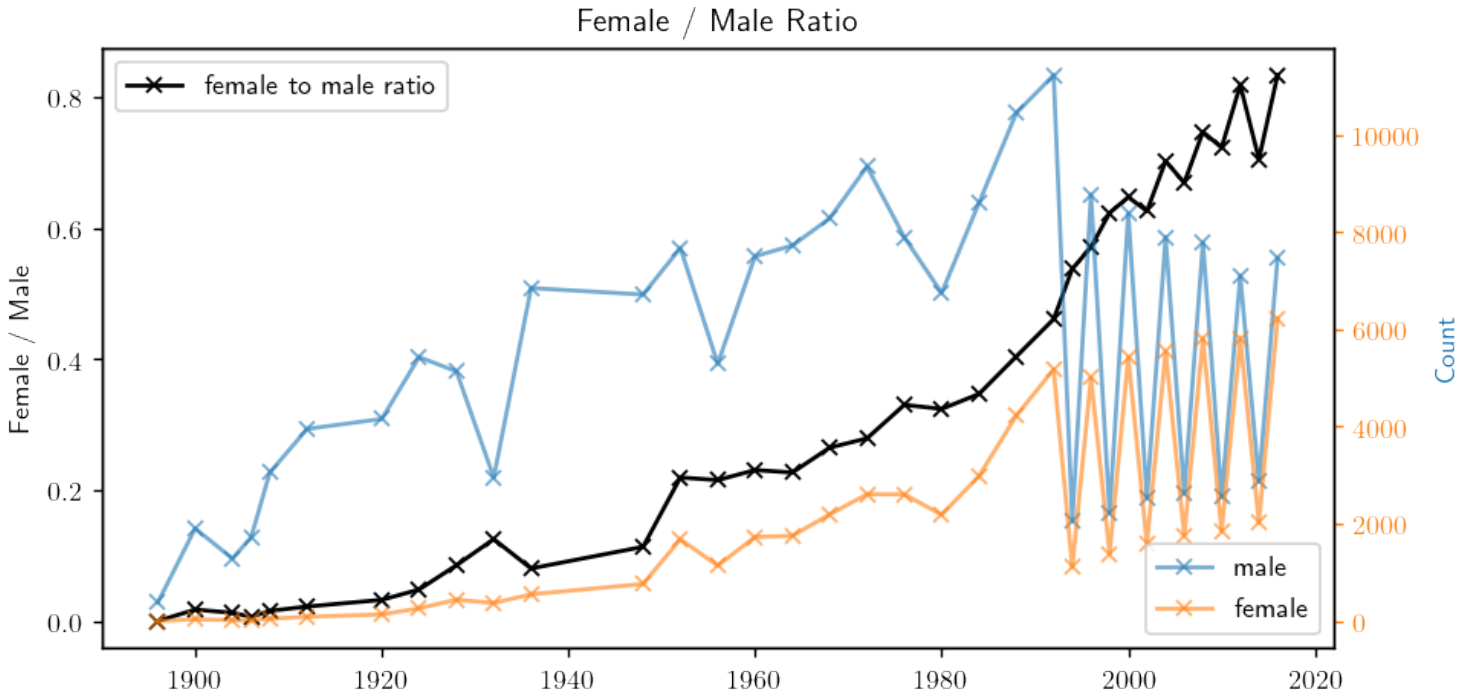


Winter games:



Male & Female Participation

Historical participation of males and females, with *female*/*male*:

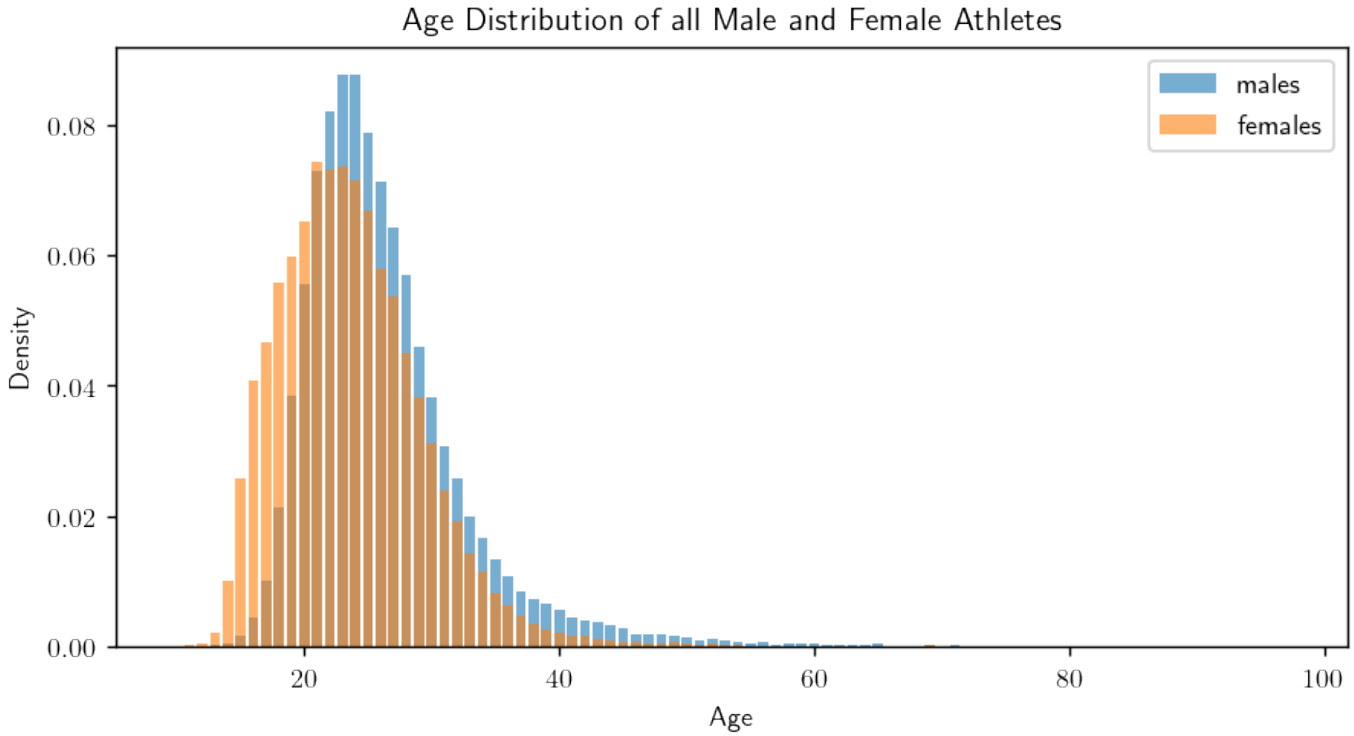


Thoughts:

The number of females per male has gradually been increasing since the Olympics started (currently at 0.8 females per male — almost 1!). There are two bumps in *female*/*male*: one between 1920 – 1940 and one between 1950 – 1960. These could have happened as a result of the wars, whereby men will have either died, or been left too injured to compete.

Age Distribution of Athletes

Age distribution of all male and female athletes since 1924:

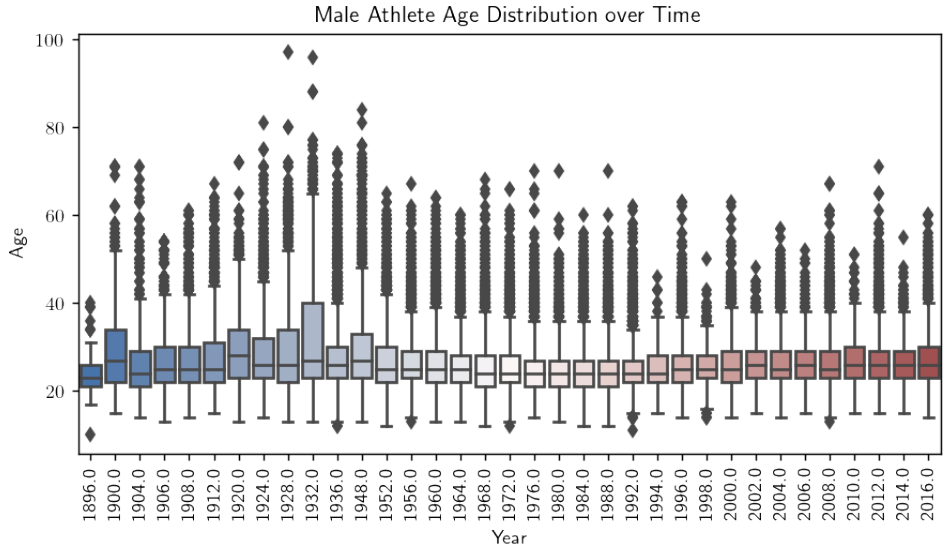


Thoughts:

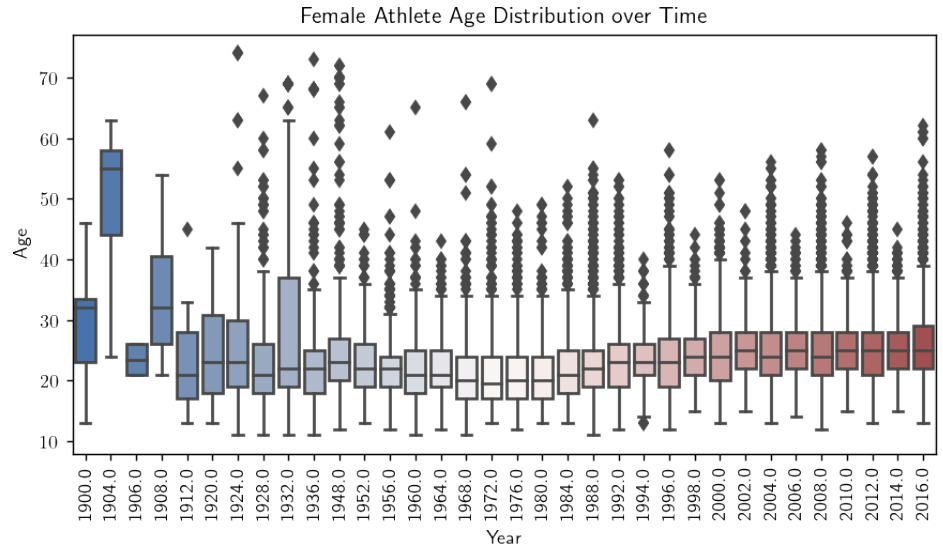
Both distributions exhibit a positive skew (so athletes are generally younger than old). With females being younger than males overall. The male distribution also has a long and thin tail in the higher ages.

(1) Female & Male Age Distributions Over Time

Males:

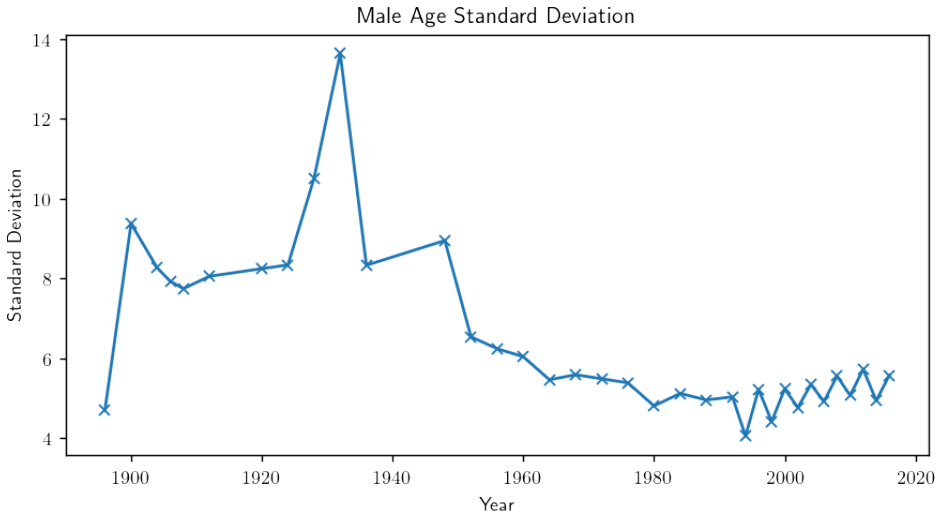


Female:

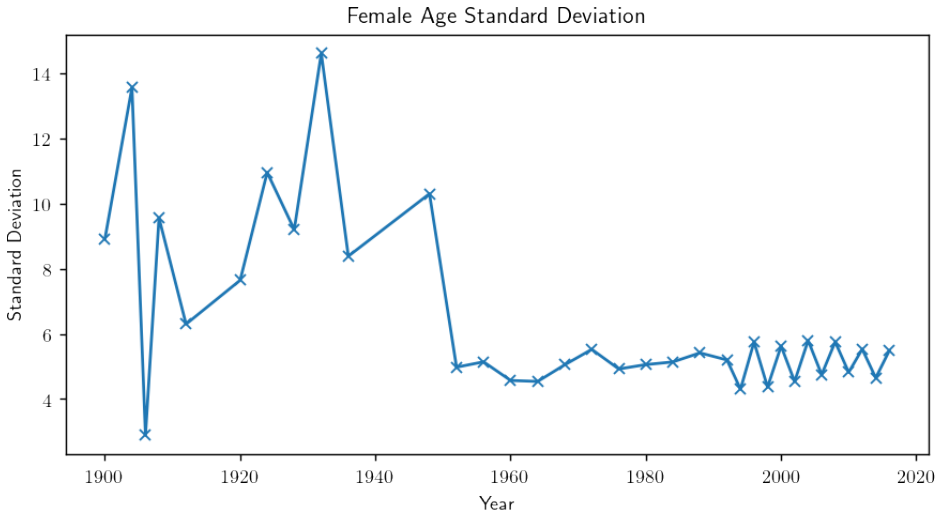


(2) Female & Male Age Distributions Over Time

Males:



Female:



(3) Female & Male Age Distributions Over Time

Results:

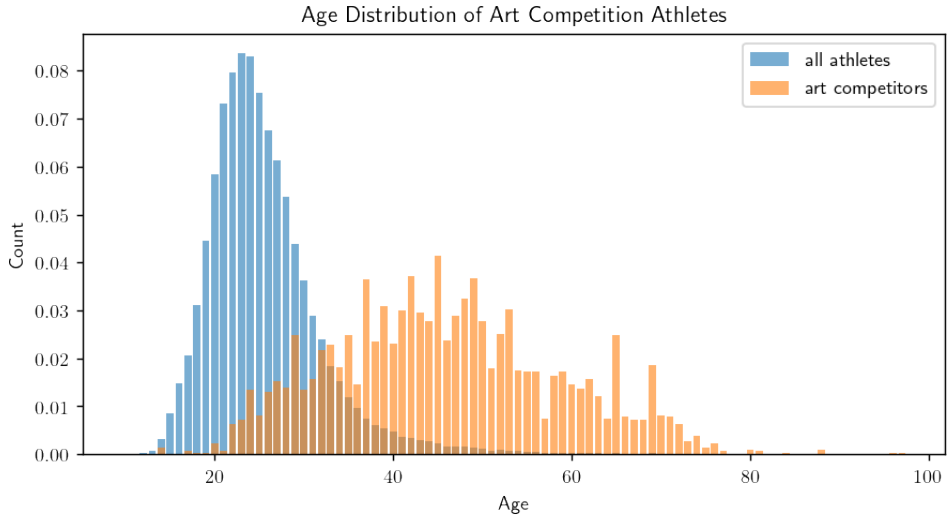
- Up until $\sim$ 1960 there is a lot of spread variation in the ages of Olympians, as demonstrated by the standard deviation plots on the previous slide. This must be caused, in part, by the missing data we identified earlier — all of which also stabilises around 1960.
- Looking at the box plots, there occurs a sudden shift towards older ages at year 1932. Why could this be? Studying the data further, it's because of the inclusion of art competitions in the olympics (result below).

```
In [99]:
1 # Looking into 1932 further, as the standard deviation is quite high for both females and males:
2 merged_df[(merged_df['Year']==1932) & (merged_df['Age']>=70)]['Sport'].value_counts()

Out[99]: Art Competitions    40
         Name: Sport, dtype: int64

In [100]:
1 merged_df[(merged_df['Year']==1928) & (merged_df['Age']>=60)]['Sport'].value_counts()

Out[100]: Art Competitions    103
          Equestrianism         2
          Fencing               1
          Name: Sport, dtype: int64
```

Let's look further into these art competitions...

Art Competitions

Normalised distribution comparing the age of art competitors and all athletes:



—i.e. they are much older.

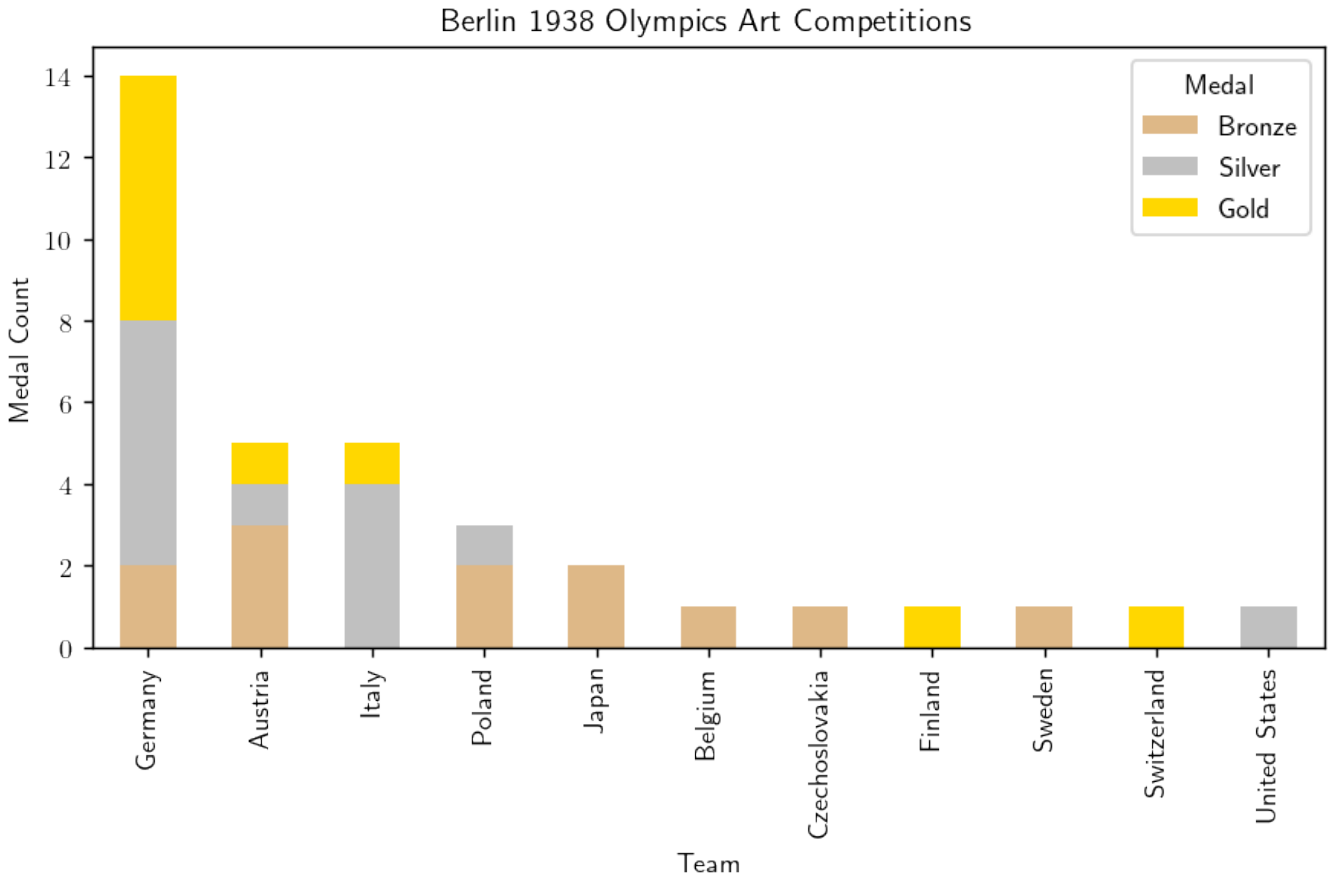
Which countries did best in the art competitions?

Out[115]:

Germany	26
France	15
Italy	14
Austria	10
Great Britain	9
Denmark	9
United States	9
Switzerland	9
Belgium	8
Poland	8
Netherlands	7
Hungary	5

So, what made Germany so successful?...

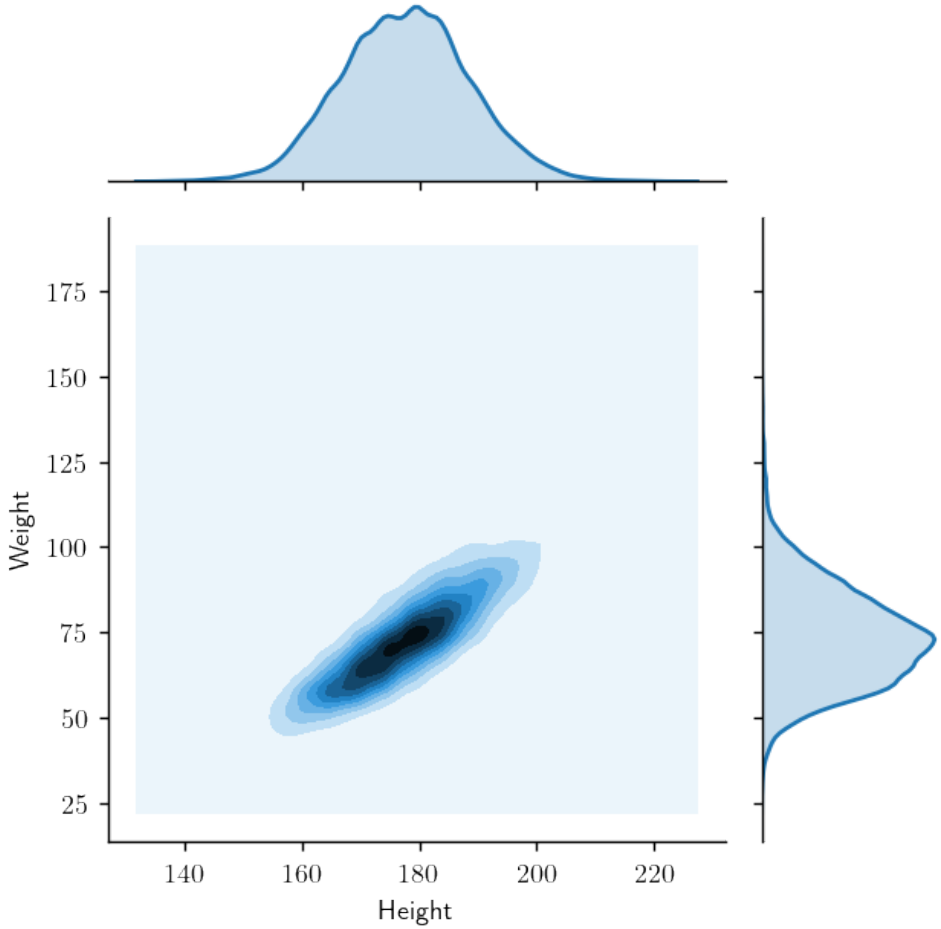
Berlin 1938



The 1938 Berlin Olympics was hosted by the Nazi regime. This is where the Germans compiled most of their success in the art competitions; as they used it as a platform for pro-Nazi propaganda. No surprise they did so well!

Athlete Height & Weight

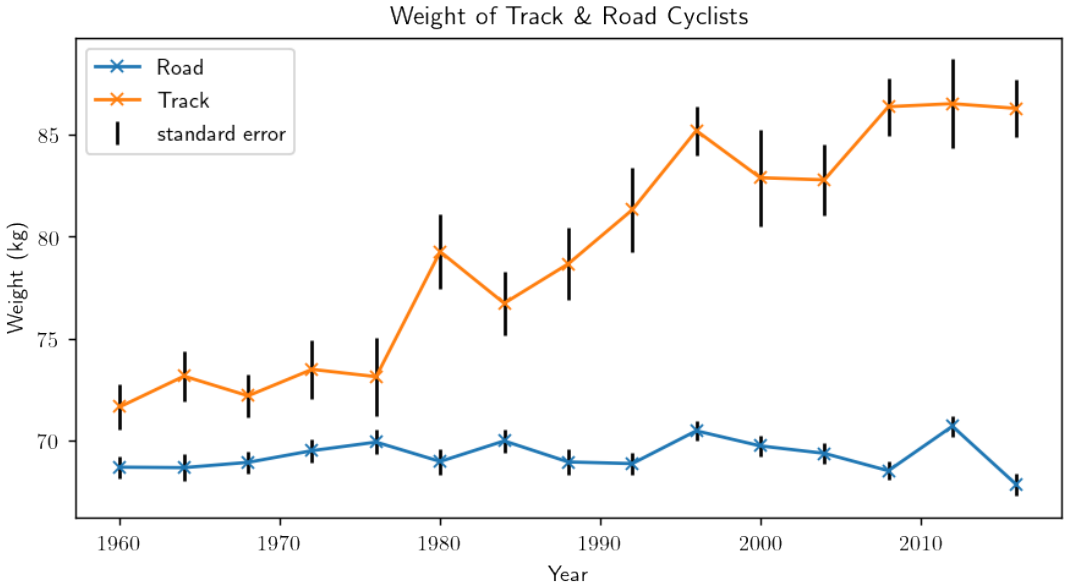
Height and Weight distribution of all athletes since 1896:



Let's look at how weight of cyclists has varied over time...

Cyclist Weight

Cyclist weight differs drastically depending on the discipline. With track cyclists generally being more heavier and powerful, and road cyclists more lighter and geared for endurance. Let's see how the average weight of Men's Road Race (road), and Men' Sprint (track) have varied over time:



Thoughts:

Unsurprisingly, road cyclists have remained relatively light. Whilst, track cyclists have gotten gradually heavier over time. The next slide compares the two physiques...

Cyclist Physiques

Track cyclist:

Road cyclist: