DATA SCIENCE CAPSTONE PROJECT

MILESTONE 4

Franco Paz- UPC

CONTENTS

- Project presentation
- Questions to Answer
- Initial Hypotheses
- Data Analysis Approach
- Technical Challenges
- Detail: Entity Relationship Diagram (ERD)
- Initial Findings
- Deeper Analysis / Going Broader
- Final Findings (Hypotheses Results)
- Recommendations

PROJECT PRESENTATION

- Client/Dataset: **SportsStats (Olympics Dataset - 120 years of data)**
  SportsStats is a sports analysis firm partnering with local news and elite personal trainers to provide "interesting" insights to help their partners.

- Objective: establish whether a **correlation** exists between the country of medal-winning Summer and Winter Games athletes and their age or their morphology over the years.

- Such informations could be of a certain interest for anyone who want to have a better understanding of such relationship.

  - *For **elite trainers***: could help to identify future talents and where they come from.

  - *From a **commercial** side*: could provide informations to sportswear designers and sellers to target specific countries, clothing sizes, etc.

QUESTIONS TO ANSWER

- Q1. Which countries were the most prolific during Summer and Winter Games over the years?
- Q2. How the height/weight ratio and age of medalists is evolving through the years?
- Q3. Is there any correlation between the country of medal-winning athletes and their age or their morphology?

INITIAL HYPOTHESES

- I am expecting to see **best results for highly populated and developed countries** such as USA, China or Russia.
- Regarding the age, talented athletes are detected earlier and earlier, which would lead to **decrease the average age of medalist**.
- For their morphology, trainings of athletes evolved a lot over years in order to optimize their performance. **Different morphologies** are expected compared to the past.
- I suppose that medalists of **many (or all) countries are within same ranges of ages or morphologies**.

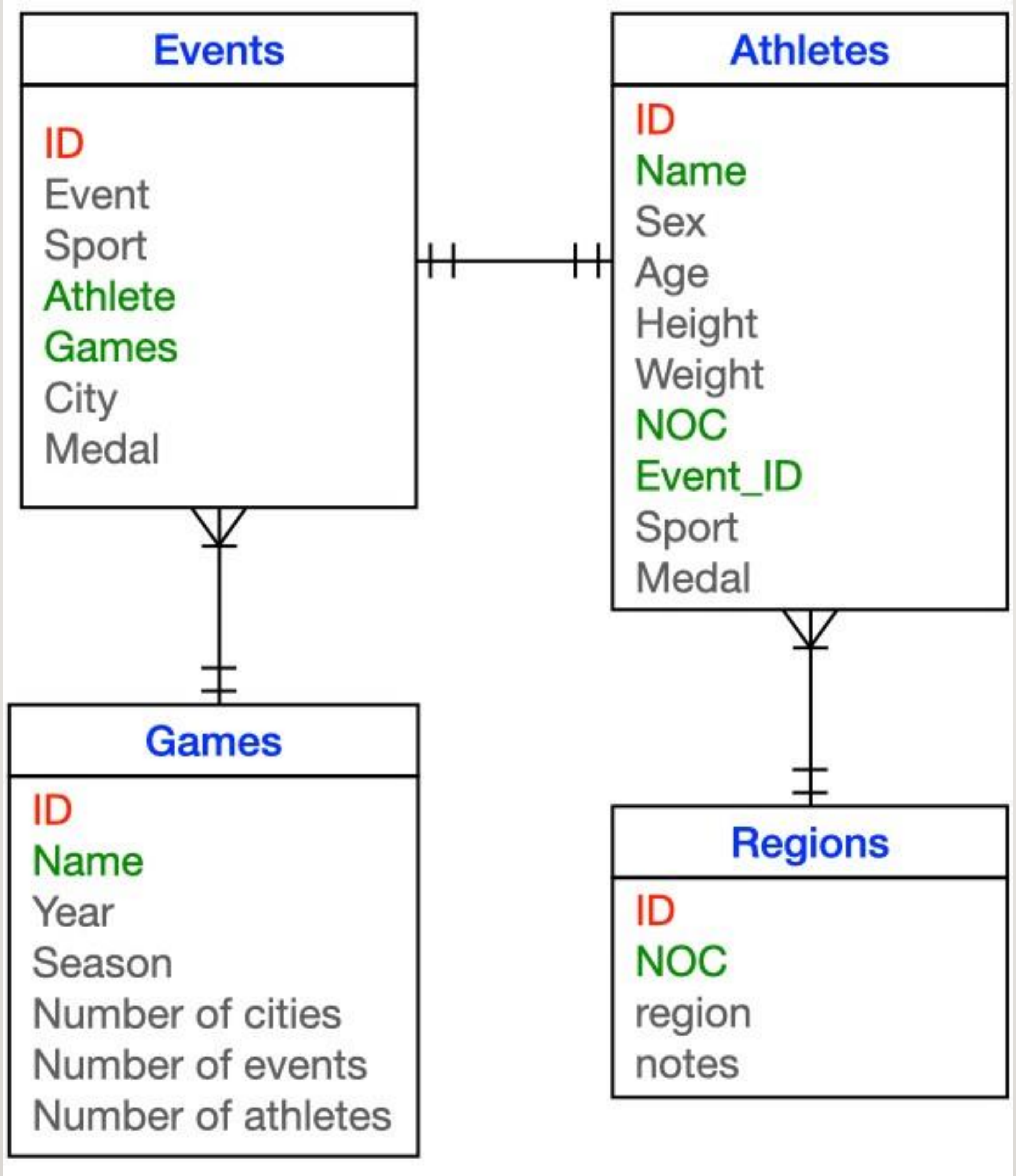
DATA ANALYSIS APPROACH

- Relationship between **country of athletes** and the **year of Winter or Summer Games**.
- The **age and height/weight ratio averages** will be also studied.
- **Statistics** will also be used by evaluating **minimum, maximum or average** of data, as well as more advanced technics such as **Pearson correlation coefficients**
- **Data visualization** with line plots, heatmaps, etc.
- <u>Tools:</u> Jupyter Notebook, SQL (Pandasql library), Python libraries

TECHNICAL CHALLENGES

- Encountered challenges with **data visualization for representing large amount of data** when establishing the relationship with countries
- Limitations of **Pandasql** Python library to execute SQL commands

ENTITY RELATIONSHIP DIAGRAM



Events

ID
Event
Sport
Athlete
Games
City
Medal

Athletes

ID
Name
Sex
Age
Height
Weight
NOC
Event_ID
Sport
Medal

Games

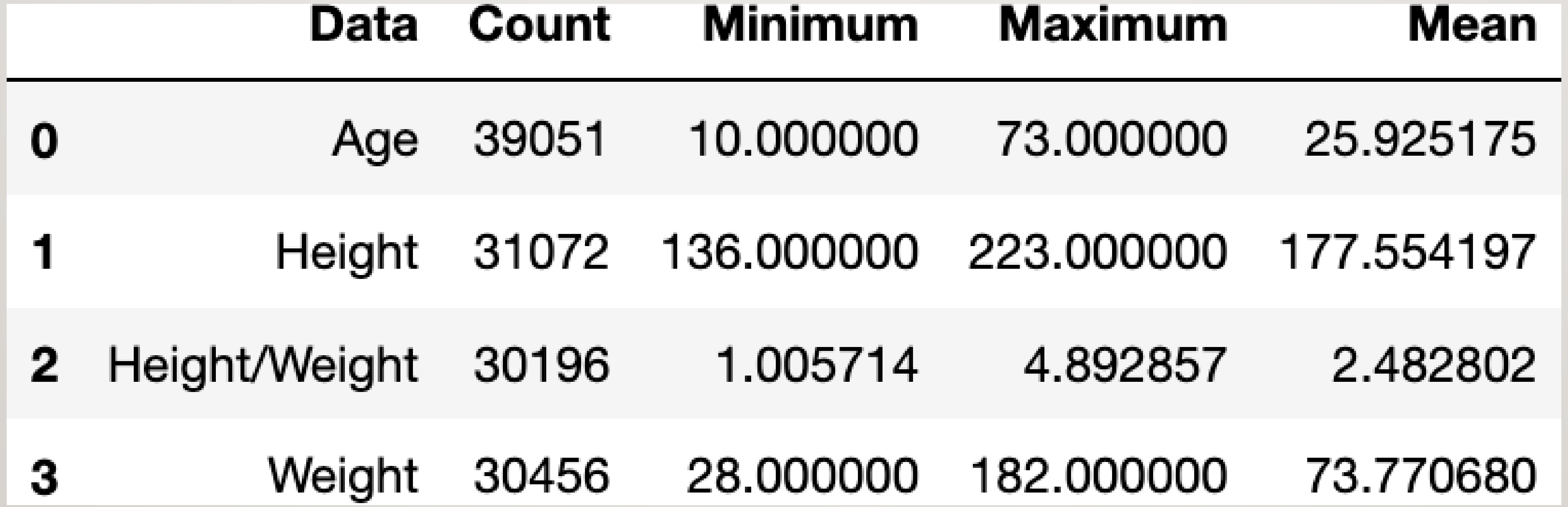
ID
Name
Year
Season
Number of cities
Number of events
Number of athletes

Regions

ID
NOC
region
notes

INITIAL FINDINGS

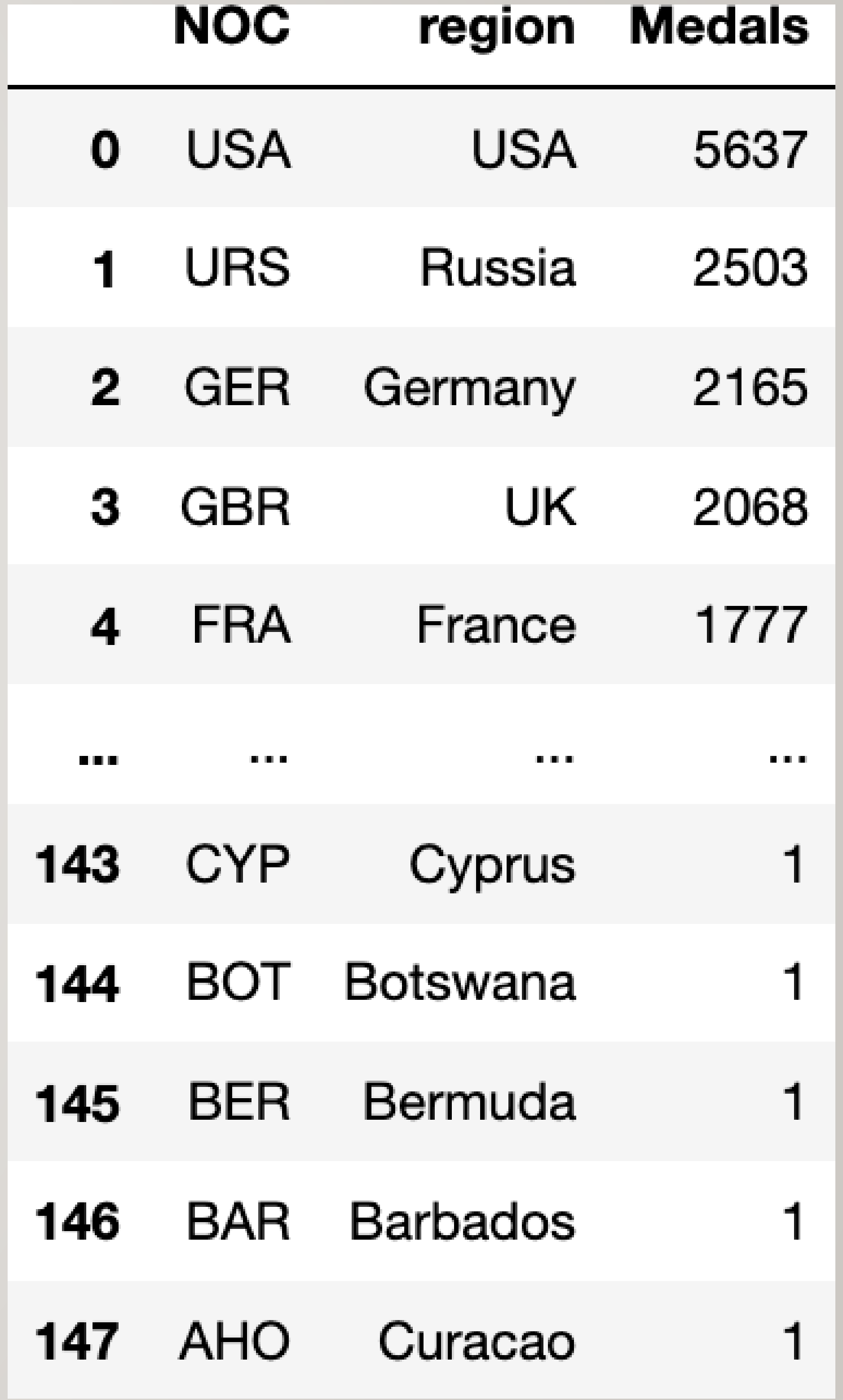
- **Contains null-data**
- In average, medalists are of **medium size** (mean 177.55) **and weight** (mean 73.77), with an average ratio of 2.48. They are also **relatively young** (mean 25.93).



	Data	Count	Minimum	Maximum	Mean
0	Age	39051	10.000000	73.000000	25.925175
1	Height	31072	136.000000	223.000000	177.554197
2	Height/Weight	30196	1.005714	4.892857	2.482802
3	Weight	30456	28.000000	182.000000	73.770680

INITIAL FINDINGS

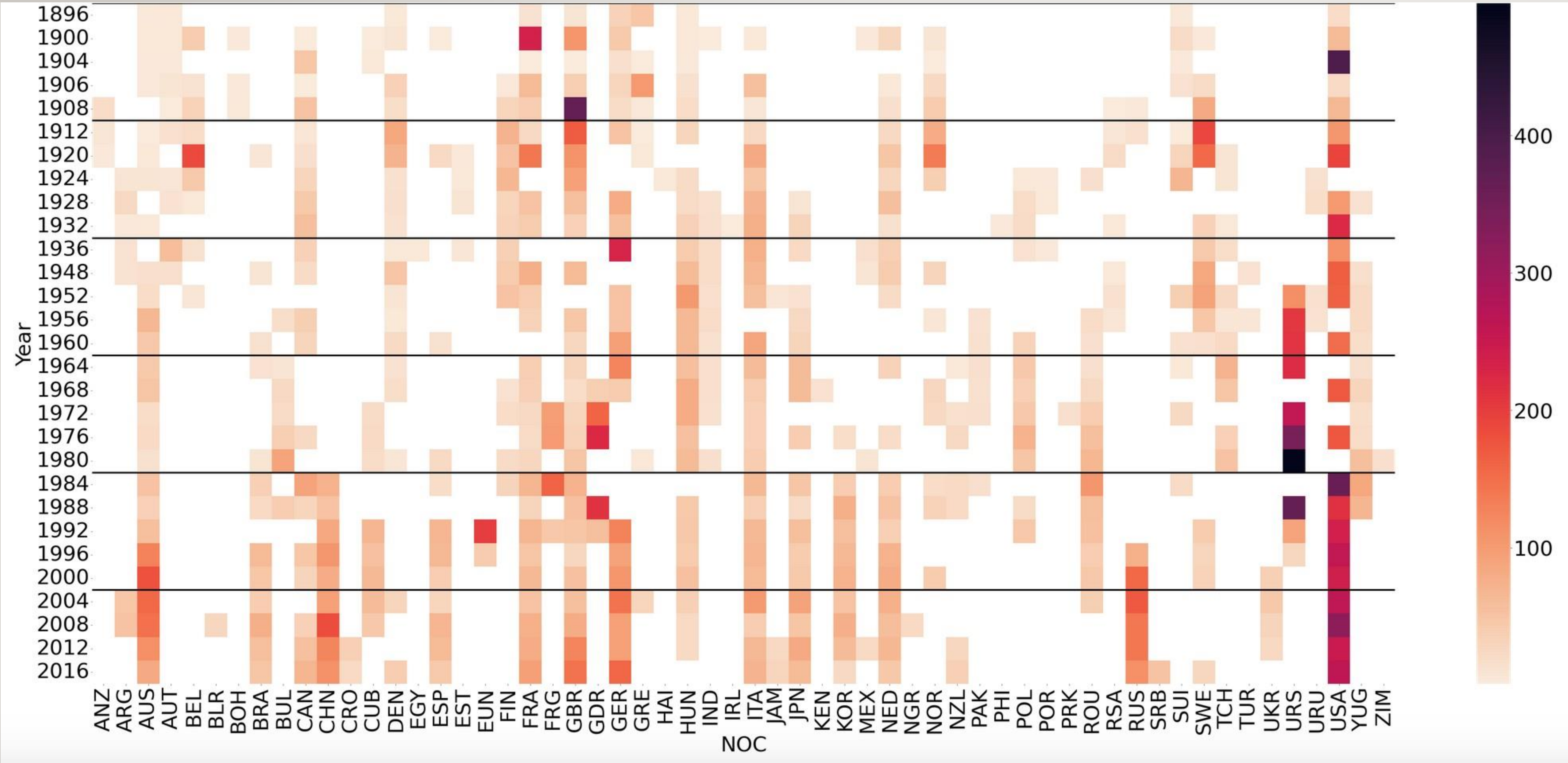
- The fifth first ranked countries are developed countries but also highly populated (so statistically a higher number of athletes)
- In contrast, the fifth last ranked countries correspond to small countries with a rather small number of citizens (so statistically a lower number of athletes)



	NOC	region	Medals
0	USA	USA	5637
1	URS	Russia	2503
2	GER	Germany	2165
3	GBR	UK	2068
4	FRA	France	1777
...	...	...	...
143	CYP	Cyprus	1
144	BOT	Botswana	1
145	BER	Bermuda	1
146	BAR	Barbados	1
147	AHO	Curacao	1

INITIAL FINDINGS

# Medals in Summer Games



- USA dominates the Summer Games for more than 40 years.

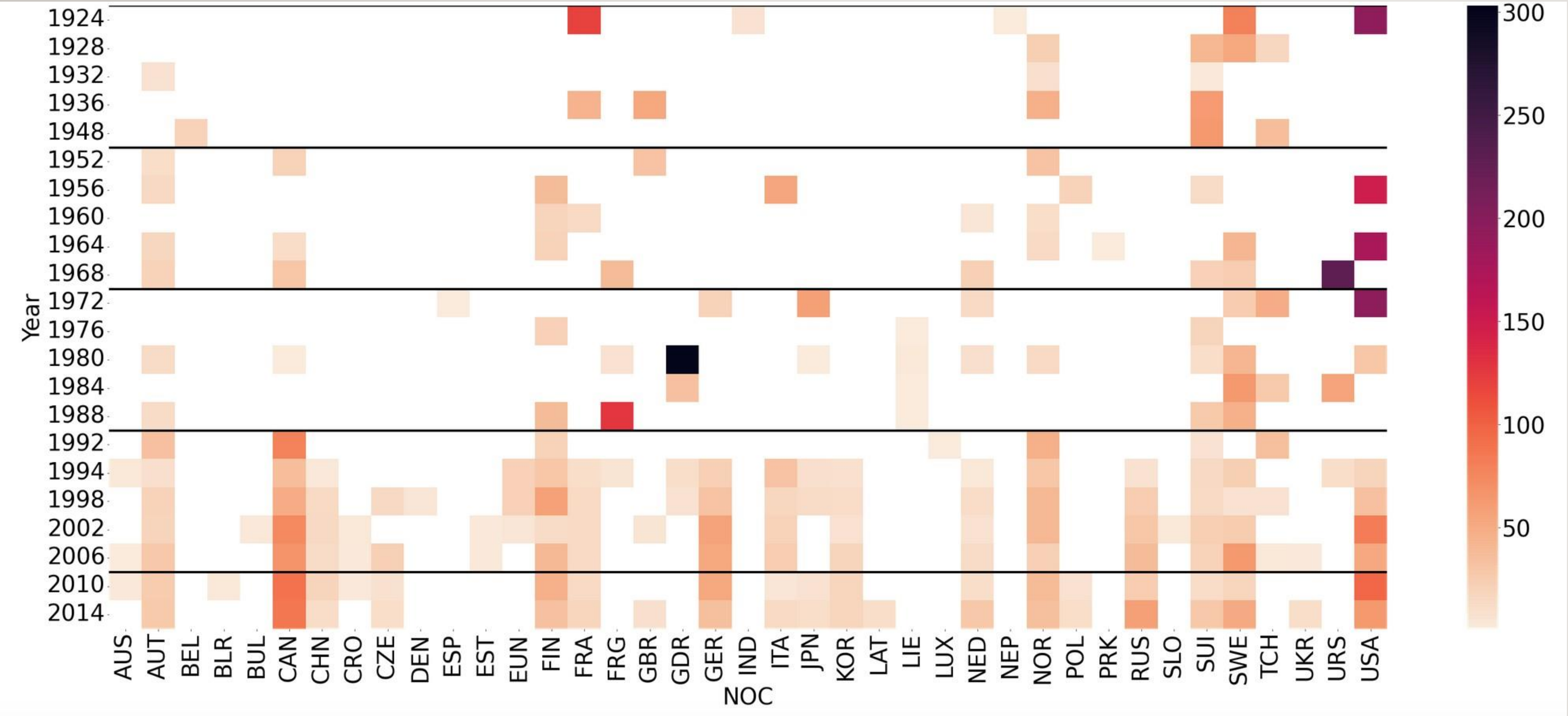
- For 20 years, USA are followed by some European nations (Germany, France, UK, Italy, etc.) as well as Asian nations (Russia, China, Japan, etc.) or Australia.

- From the 50s to 70s, these Games were dominated by ex-URSS.

✓ *1st Hypothesis*

INITIAL FINDINGS

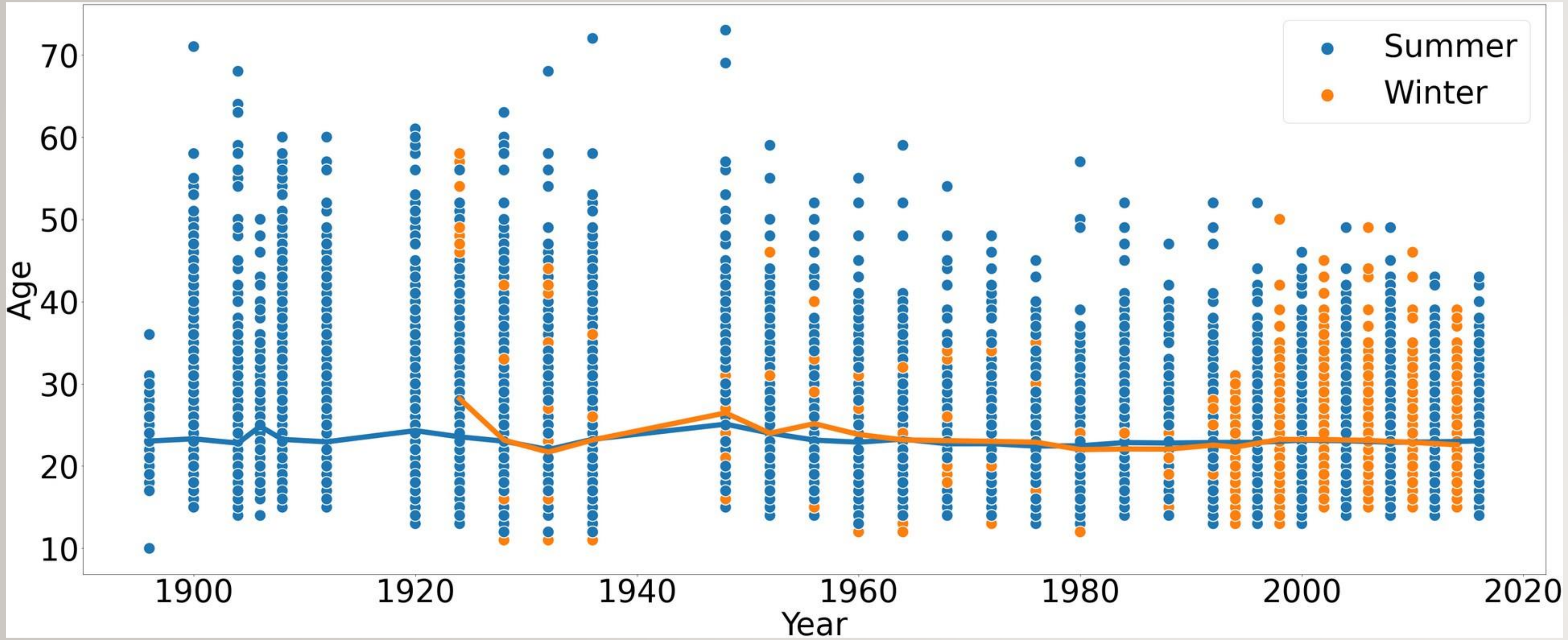
# Medals in Winter Games

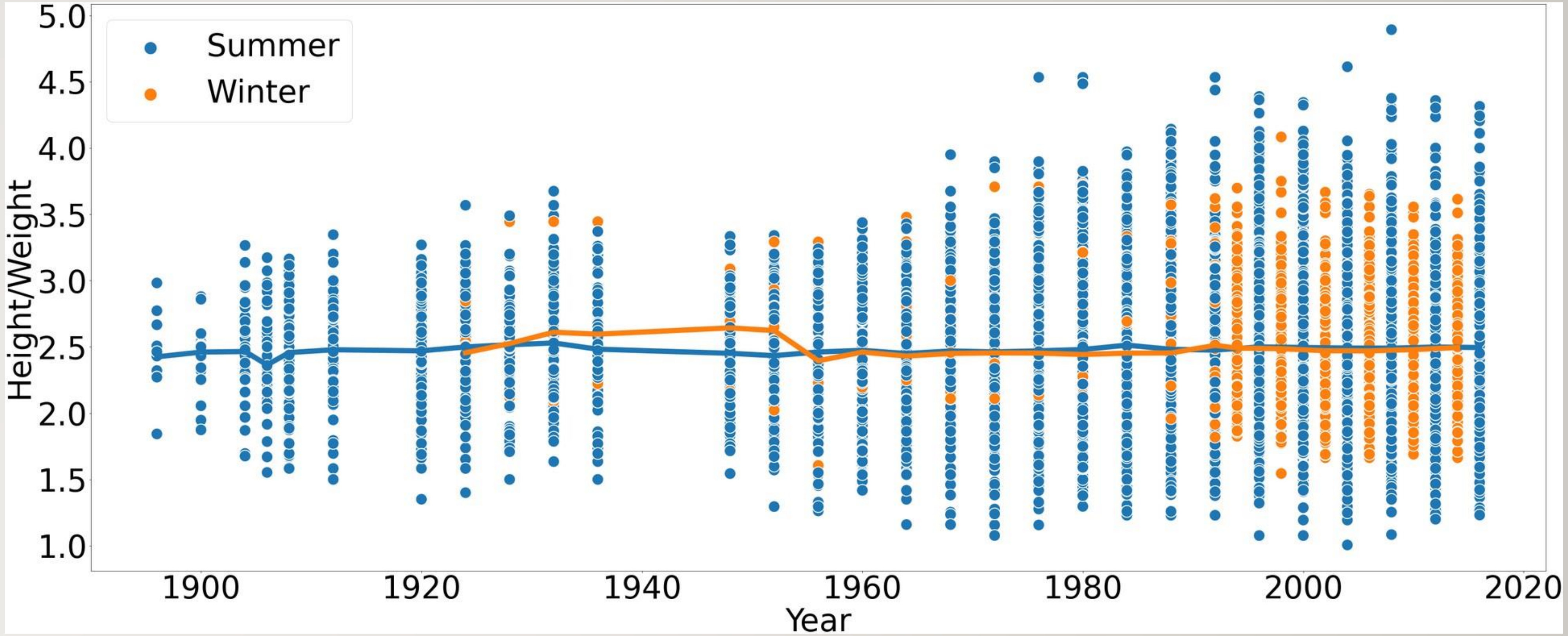


- USA also mainly dominates over the years, followed by Canada since 1992.
- Again, Russia (and ex-URSS) shown some medals, but most of time less important than Nordic countries such as Sweden, Norway or Finland.

✓ *1st Hypothesis*

INITIAL FINDINGS





- Mean age of medalists is near 25, while the distribution of data is wider for oldest years than closest years.

- Could be interesting to correlate it with the nature of sports.

✗*2nd and 3rd Hypotheses*

- Mean appears more or less constant over the years and whatever the season is, around 2.5 which is a "regular" ratio.

- In contrast, the distribution of data is wider for closest years than oldest years.

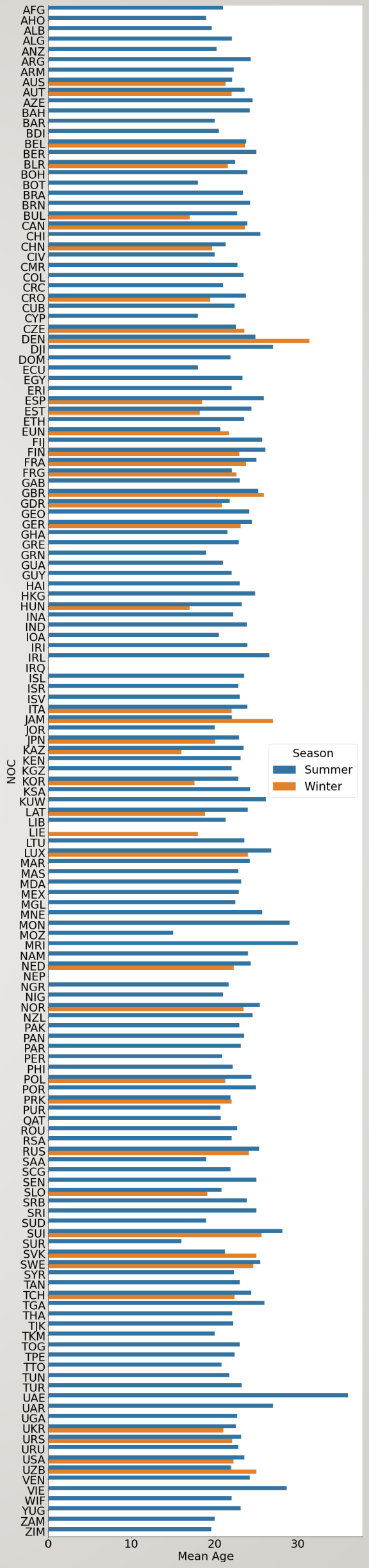
- Could be interesting to plot separately Height and Weight to get deeper insights.

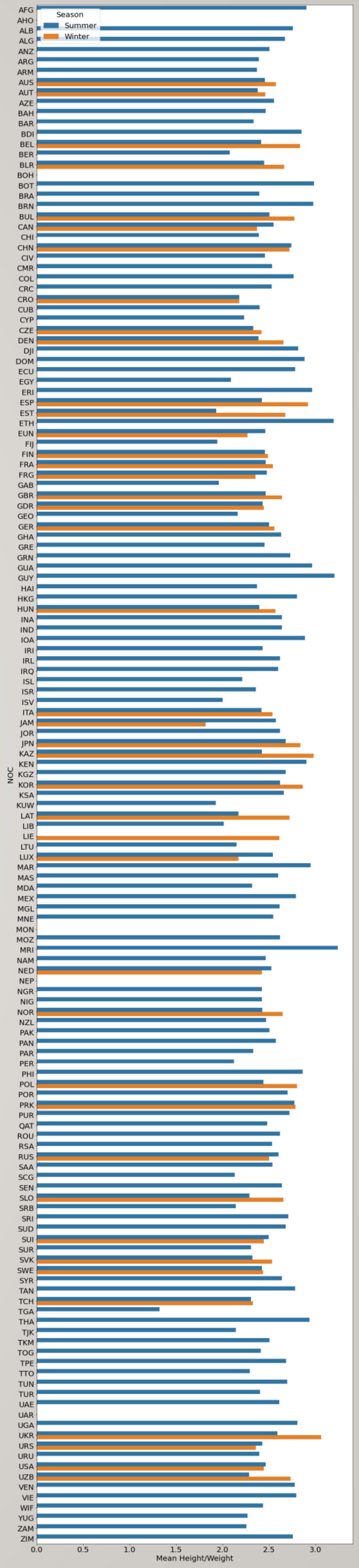
INITIAL FINDINGS

- Most of the countries show **same behavior** with medalists ranging between 20 and 25 years old, whatever the season is.

- Most of countries show **same behavior** with medalists with a mean Height/Weight ratio arount 2.3-2.5, whatever the season is.

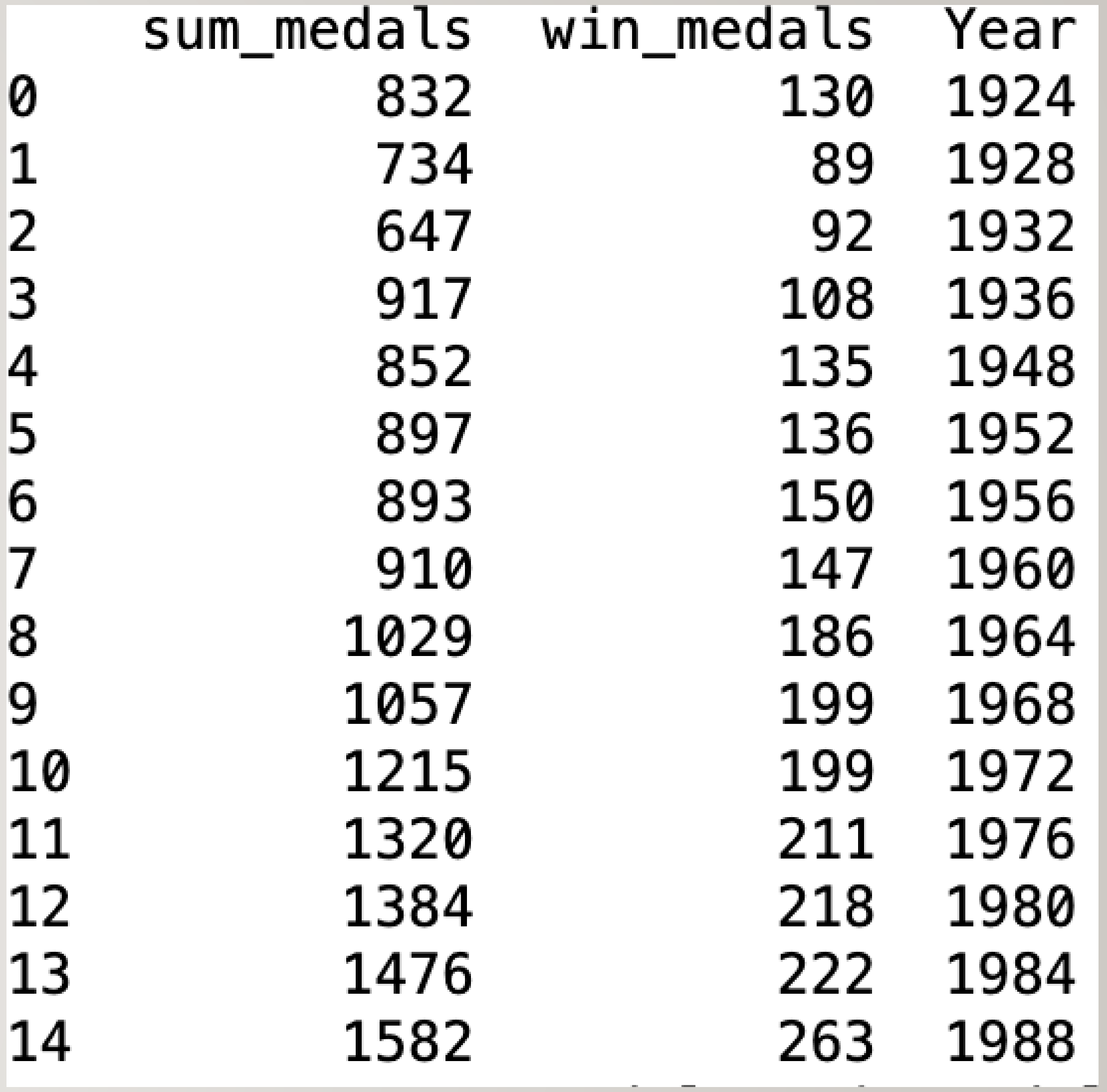
✓ *4th Hypothesis*



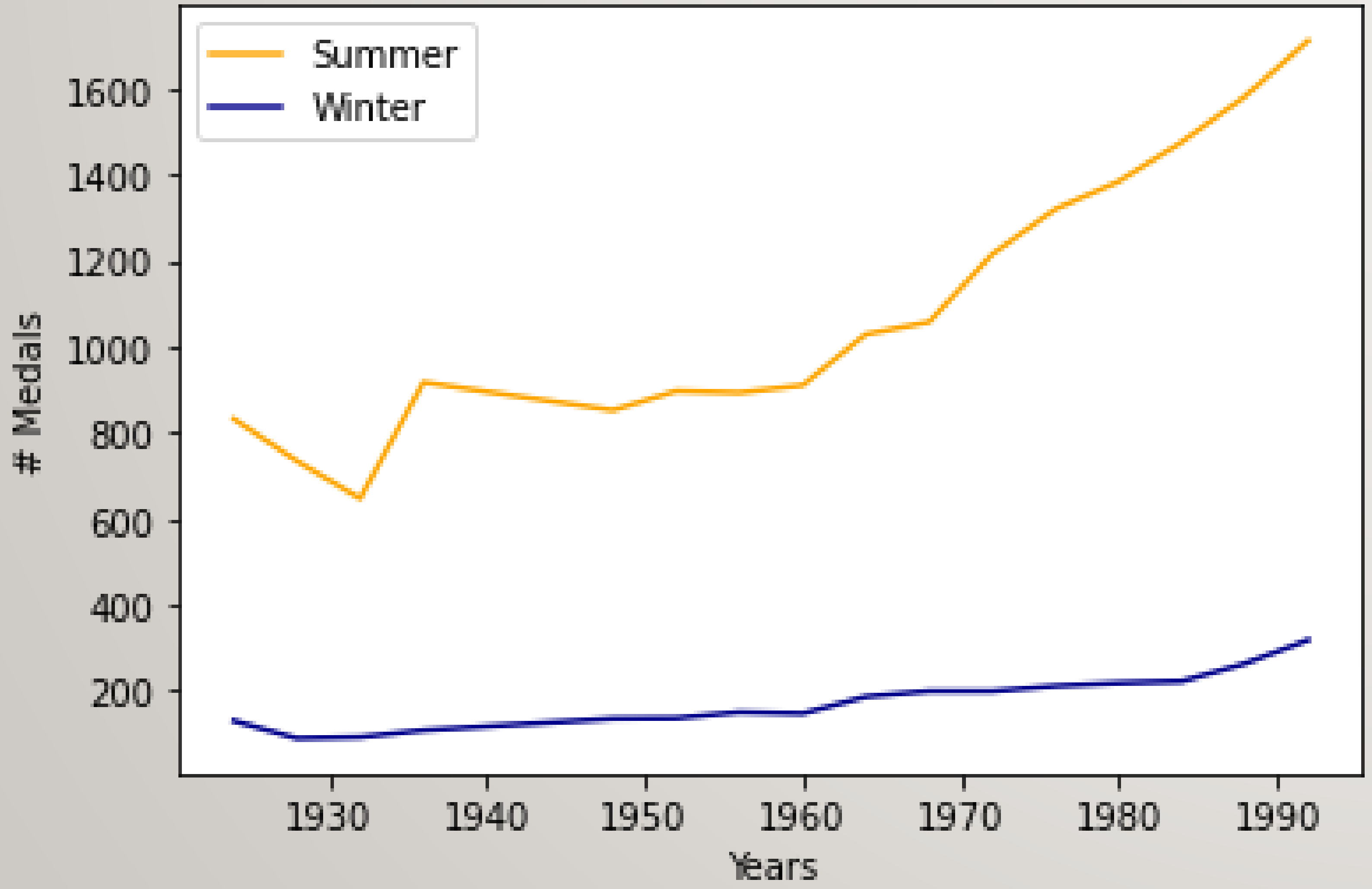


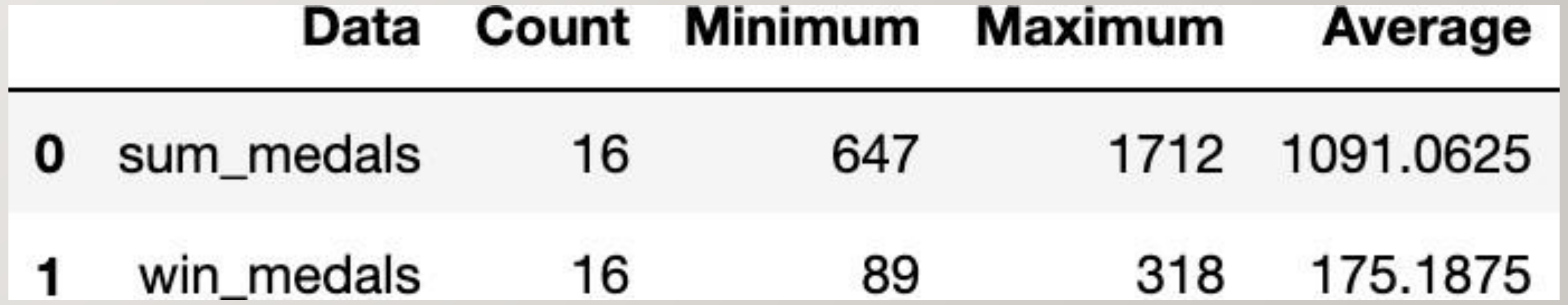
DEEPER ANALYSIS

- Correlation between the total number of medalists in the Winter and Summer Games.
- Pearson corr. coef. = 0.960
- High correlation of performances in both seasons



```
    sum_medals  win_medals  Year
0          832         130  1924
1          734          89  1928
2          647          92  1932
3          917         108  1936
4          852         135  1948
5          897         136  1952
6          893         150  1956
7          910         147  1960
8         1029         186  1964
9         1057         199  1968
10        1215         199  1972
11        1320         211  1976
12        1384         218  1980
13        1476         222  1984
14        1582         263  1988
```

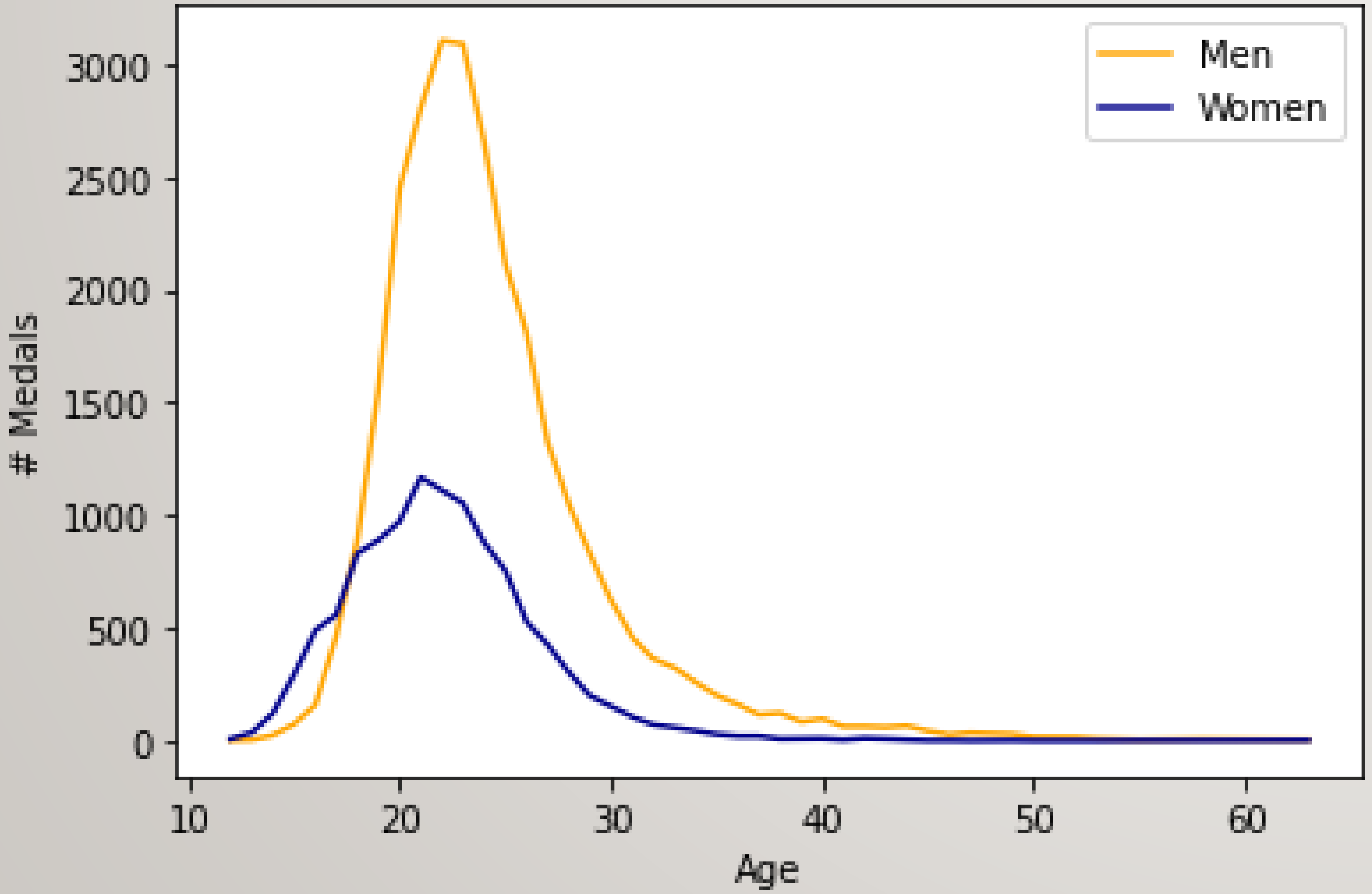


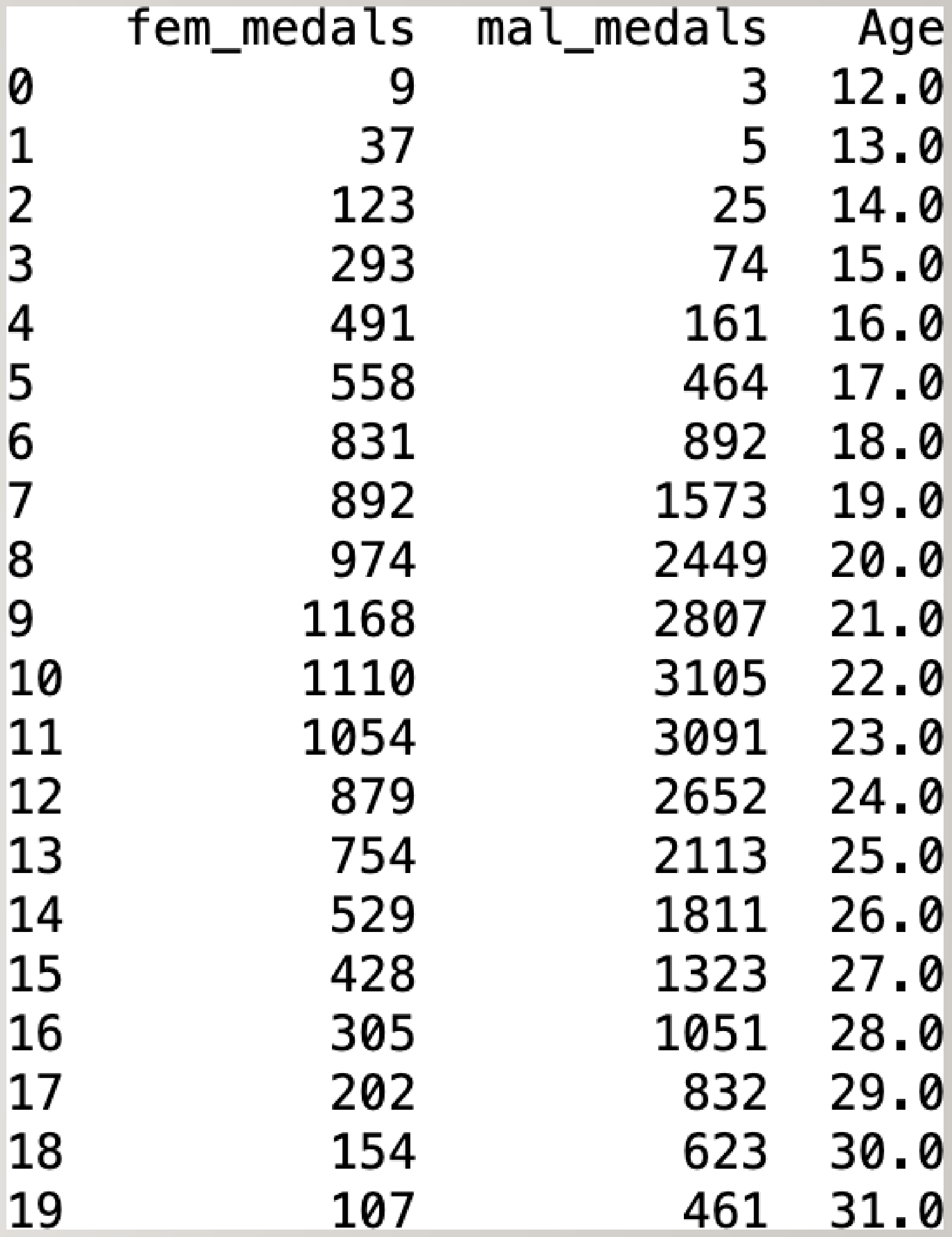


	Data	Count	Minimum	Maximum	Average
0	sum_medals	16	647	1712	1091.0625
1	win_medals	16	89	318	175.1875

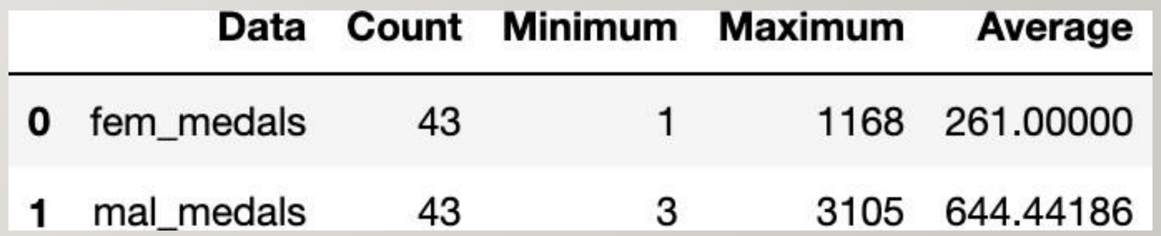
GOING ABROAD

- Correlation between the total number of women and men medalists for given ages
- Pearson corr. coef. = 0.924
- Good correlation of performances for both sexes





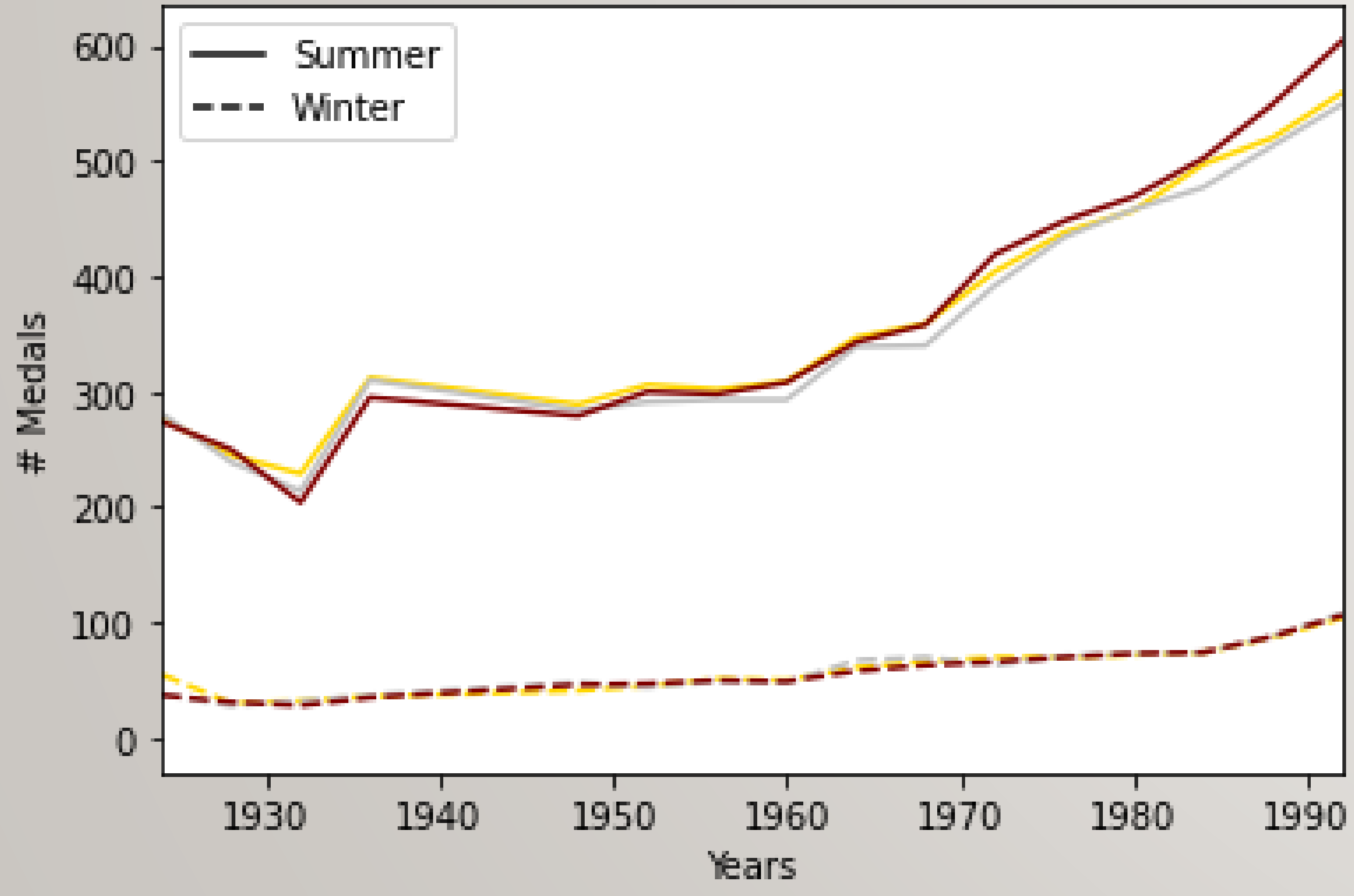
```
    fem_medals  mal_medals   Age
0            9           3  12.0
1           37           5  13.0
2          123          25  14.0
3          293          74  15.0
4          491         161  16.0
5          558         464  17.0
6          831         892  18.0
7          892        1573  19.0
8          974        2449  20.0
9         1168        2807  21.0
10        1110        3105  22.0
11        1054        3091  23.0
12         879        2652  24.0
13         754        2113  25.0
14         529        1811  26.0
15         428        1323  27.0
16         305        1051  28.0
17         202         832  29.0
18         154         623  30.0
19         107         461  31.0
```



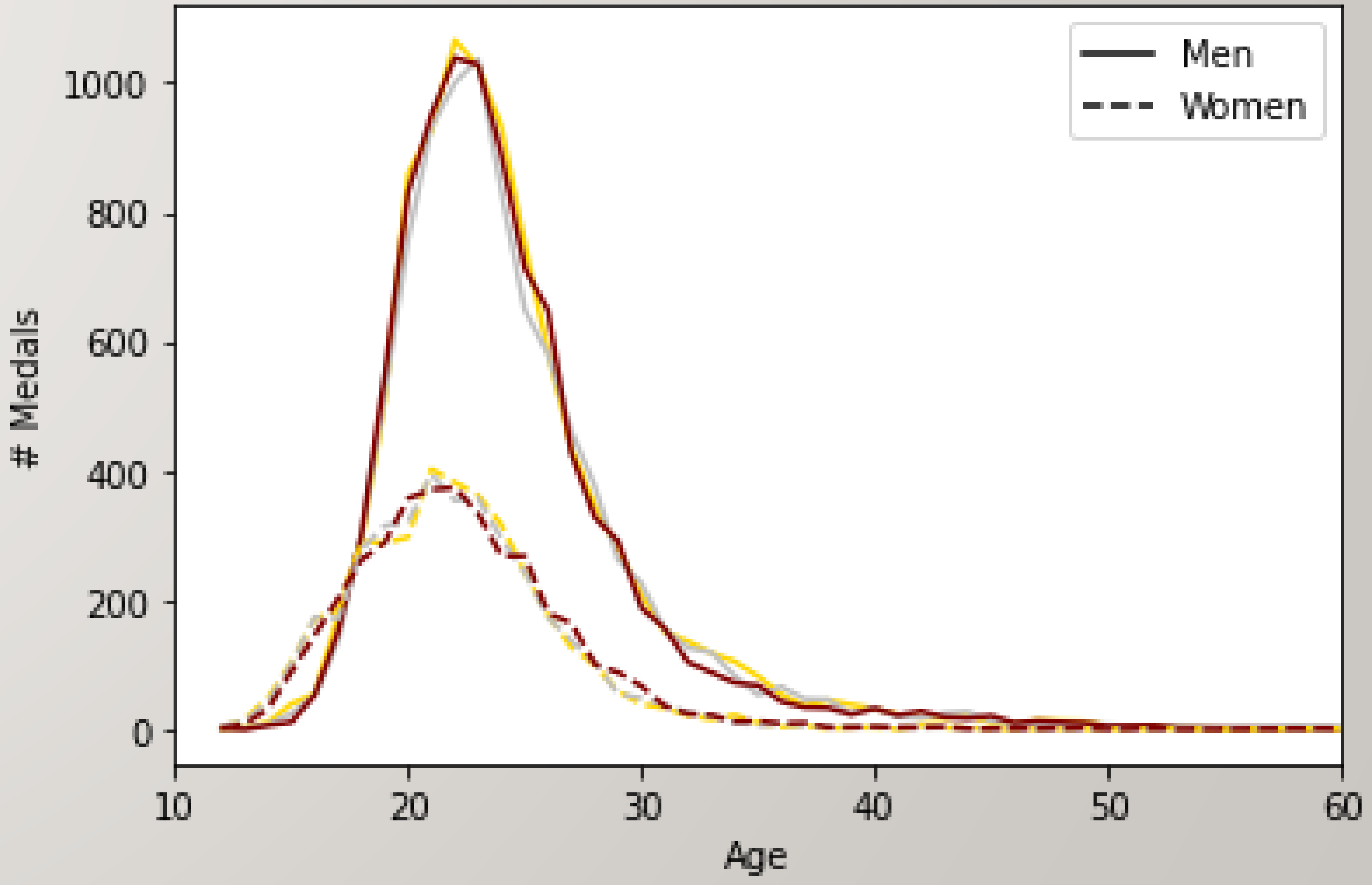
	Data	Count	Minimum	Maximum	Average
0	fem_medals	43	1	1168	261.00000
1	mal_medals	43	3	3105	644.44186

DEEPER ANALYSIS / GOING BROADER

- New metric: total number of bronze, silver and gold medals; allows to **more finely inspect similarities** in both situations, because of the **introduction of the rank**.
- Confirms high correlation in both cases



Pearson corr. coef.	*Summer vs Winter / Years*	*Men vs Women / Age*
Gold medals	0.943	0.918
Silver medals	0.937	0.917
Bronze medals	0.973	0.930



FINAL FINDINGS (RESULTS OF HYPOTHESES)

- ✓**Best results for highly populated and developed countries** such as USA, China or Russia.
- ✗The average age of medalist **is constant** over the years
- ✗The average of morphologies **is constant** over the years
- ✓**Most of countries** have medalists within same range of ages or morphologies**.**

RECOMMENDATIONS

- The average age and morphology of medalists being the same between countries, I would suggest to sportswear sellers that clothes do not have to present specific sizes or mensurations depending on the country.

- Nordic countries show logically good performances during Winter Games. I would advise concerned elite trainers to take more informations on those countries (their technics, etc.).

- To complete this analysis, it would be interesting to go deeper by **distinguishing categories of sports**. That could enable to find new correlations (more specifically with the morphology of medalists).