Olympic Games Data Analysis
SQL for Data Science Capstone Project

Mariana Ermano
January 2023

Contents

Description

Questions to Answer / Hypotheses / Approach

Technical Challenges

About the Dataset

Detail: Entity Relationship Diagram (ERD)

Initial Findings

Deeper Analysis

Hypotheses Results

Description

I want to analyze if there are any specific **factors** that may contribute to **winning medals** in the **Olympic Games**.

These insights may be helpful to try and **predict** the winning probability for a specific country/team.

This project may also appeal to sport enthusiasts looking for interesting facts about the Olympics.

The analysis will be focused on whether there is an impact on the number of **athletes** and **medals** won by a country when **hosting** the Olympics, and the relation between these two factors.

Questions to Answer

- Do the Olympics **host countries** have a higher number of **athletes**? Would this contribute to winning **more medals**?
- Do **men** win more medals than **women**?
- What is the **sport category** with the highest medals-athletes ratio?

Initial Hypotheses

- **Hosting** the Olympics correlates to a **higher** number of **athletes** participating in the events, that may contribute to winning **more medals**.
- In proportion to the number of athletes, **women** win **more medals** than men.

Data Analysis Approach

- I will be primarily looking at relationships between **frequency** metrics, like medal and athlete counts
- The **Pearson Correlation Coefficient** will be calculated to determine if there is a correlation between metrics
- I will also calculate the **proportion** of a metric when compared to the total number (i.e.: medal-athlete ratio)
- The **ABBA tool** will be used to analyze the **statistical significance** of the results for the Host vs No Host hypothesis

Technical Challenges

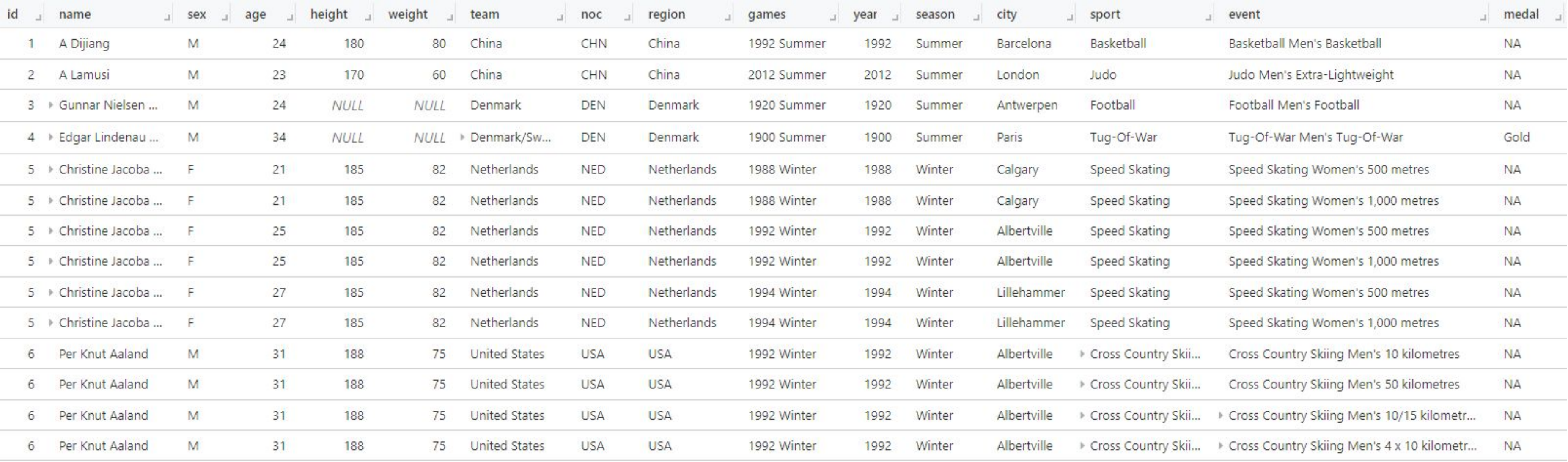
This was the first in-depth analysis I have performed with SQL, as well as my first time working with Databricks ETL tool independently. So initially it was challenging to import data and create tables.

Since this capstone project is the culmination of a SQL for Data Science specialization, I decided to transform and analyze the data using SQL only, as opposed to adding some Python code (that for big data would have made all the process easier), as it was done in the example shown on this course.

This project has definitely helped me put my SQL skills to practice and it was a great analysis exercise.

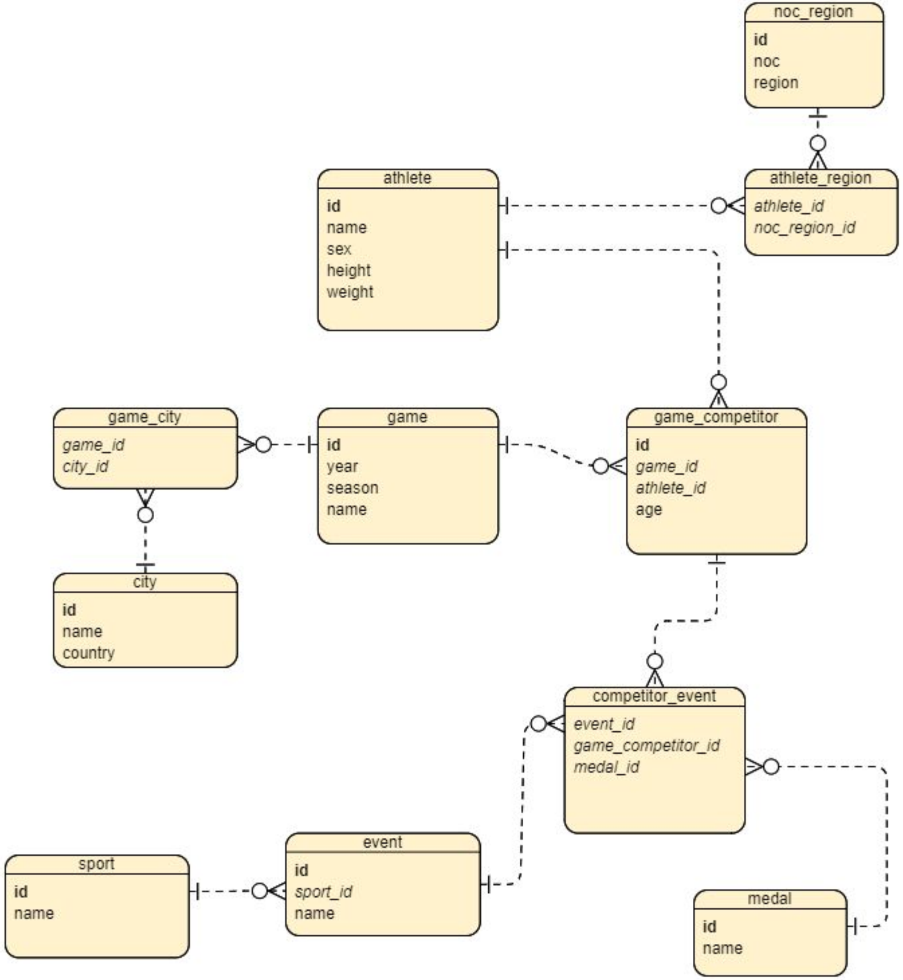
About the Dataset

It contains around **260,000** separate results for all **competitors** (not only medal winners) at all Olympic Games **events** from 1896 until 2016, including both summer and winter Olympics.



id	name	sex	age	height	weight	team	noc	region	games	year	season	city	sport	event	medal
1	A Dijiang	M	24	180	80	China	CHN	China	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NA
2	A Lamusi	M	23	170	60	China	CHN	China	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NA
3	Gunnar Nielsen ...	M	24	NULL	NULL	Denmark	DEN	Denmark	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NA
4	Edgar Lindenau ...	M	34	NULL	NULL	Denmark/Sw...	DEN	Denmark	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
5	Christine Jacoba ...	F	21	185	82	Netherlands	NED	Netherlands	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NA
5	Christine Jacoba ...	F	21	185	82	Netherlands	NED	Netherlands	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 1,000 metres	NA
5	Christine Jacoba ...	F	25	185	82	Netherlands	NED	Netherlands	1992 Winter	1992	Winter	Albertville	Speed Skating	Speed Skating Women's 500 metres	NA
5	Christine Jacoba ...	F	25	185	82	Netherlands	NED	Netherlands	1992 Winter	1992	Winter	Albertville	Speed Skating	Speed Skating Women's 1,000 metres	NA
5	Christine Jacoba ...	F	27	185	82	Netherlands	NED	Netherlands	1994 Winter	1994	Winter	Lillehammer	Speed Skating	Speed Skating Women's 500 metres	NA
5	Christine Jacoba ...	F	27	185	82	Netherlands	NED	Netherlands	1994 Winter	1994	Winter	Lillehammer	Speed Skating	Speed Skating Women's 1,000 metres	NA
6	Per Knut Aaland	M	31	188	75	United States	USA	USA	1992 Winter	1992	Winter	Albertville	Cross Country Skii...	Cross Country Skiing Men's 10 kilometres	NA
6	Per Knut Aaland	M	31	188	75	United States	USA	USA	1992 Winter	1992	Winter	Albertville	Cross Country Skii...	Cross Country Skiing Men's 50 kilometres	NA
6	Per Knut Aaland	M	31	188	75	United States	USA	USA	1992 Winter	1992	Winter	Albertville	Cross Country Skii...	Cross Country Skiing Men's 10/15 kilometr...	NA
6	Per Knut Aaland	M	31	188	75	United States	USA	USA	1992 Winter	1992	Winter	Albertville	Cross Country Skii...	Cross Country Skiing Men's 4 x 10 kilometr...	NA

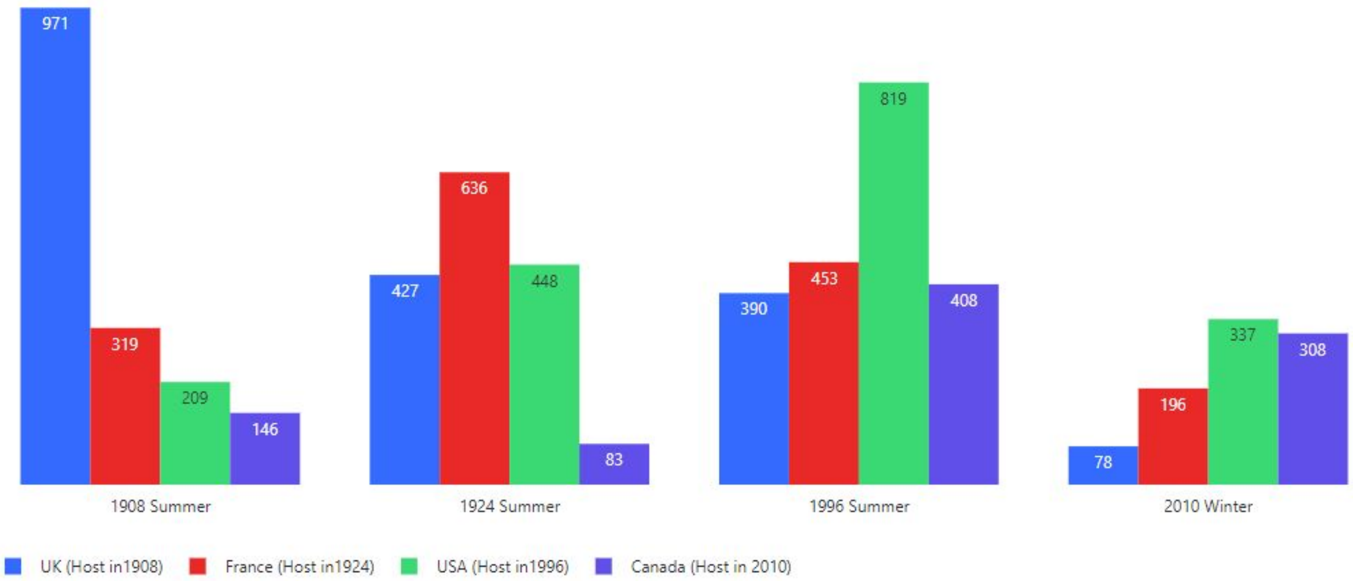
Entity Relationship Diagram



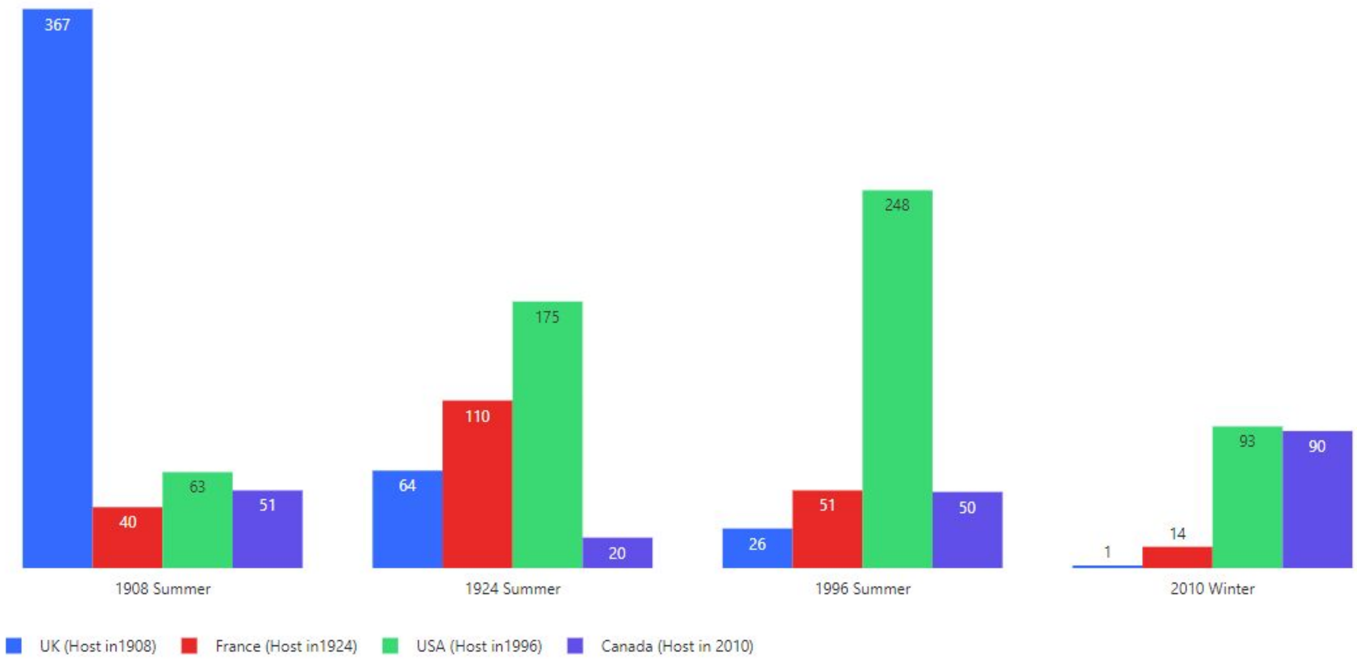
Initial Findings

Initial Findings: Part 1

In my initial exploration, just by looking at the number of **athletes** by country for a random game, I noticed that the **hosting country** was consistently high in the ranking.



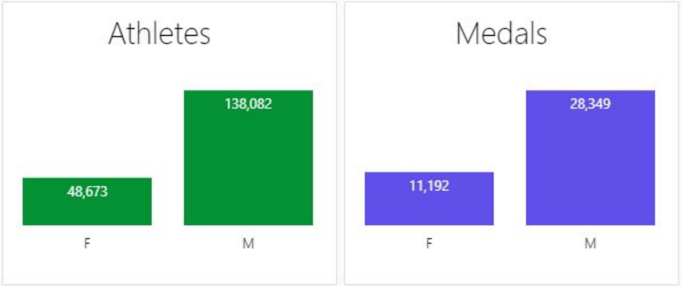
I found the same trend for the number of **medals** won by country when hosting:



Initial Findings: Part 2

The big difference between the number of athletes by sex was striking to me, so I wanted to compare the number of medals won by sex.

Below is the result of my initial exploration:



While exploring the data, I realized that because almost all of the events would have the custom of the sequence of gold, silver, and bronze for the first three places (some sports award two bronze medals per competition), the number of medals won by each sex was rather determined by the number of events for each type.

Men and **women** would not **compete** against each other, except on the **mixed events**. For this reason, I decided to analyze the **medals by athlete ratio** by sex for mixed events only.

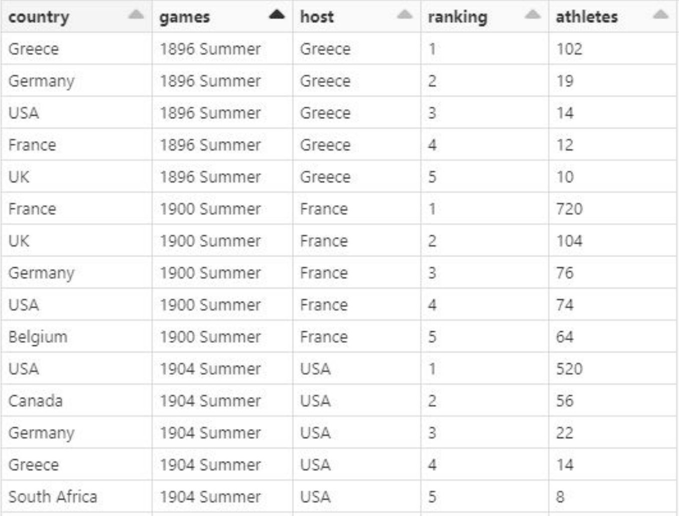
Deeper Analysis

Analyzing athletes and medals of host countries

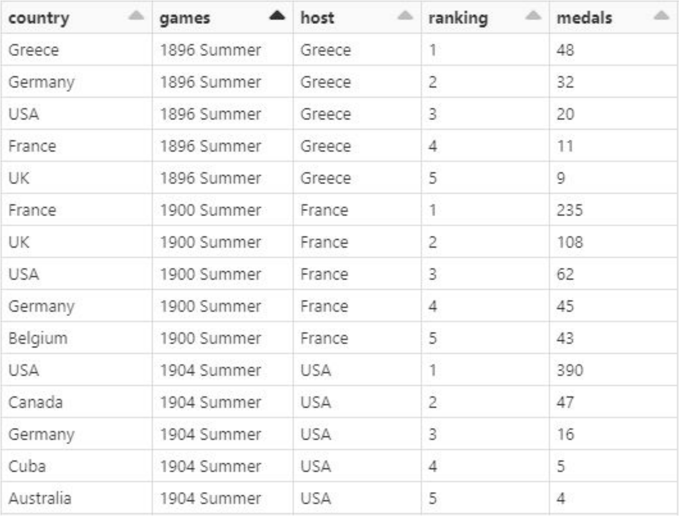
To conduct this analysis, I created a table that lists the **top 5 countries** with higher number of **athletes** by game, and the **host country** of that game. Then did the same for the number of **medals**.

There is a total of **25 host countries** and **52 Olympic Games** (1896 - 2016).

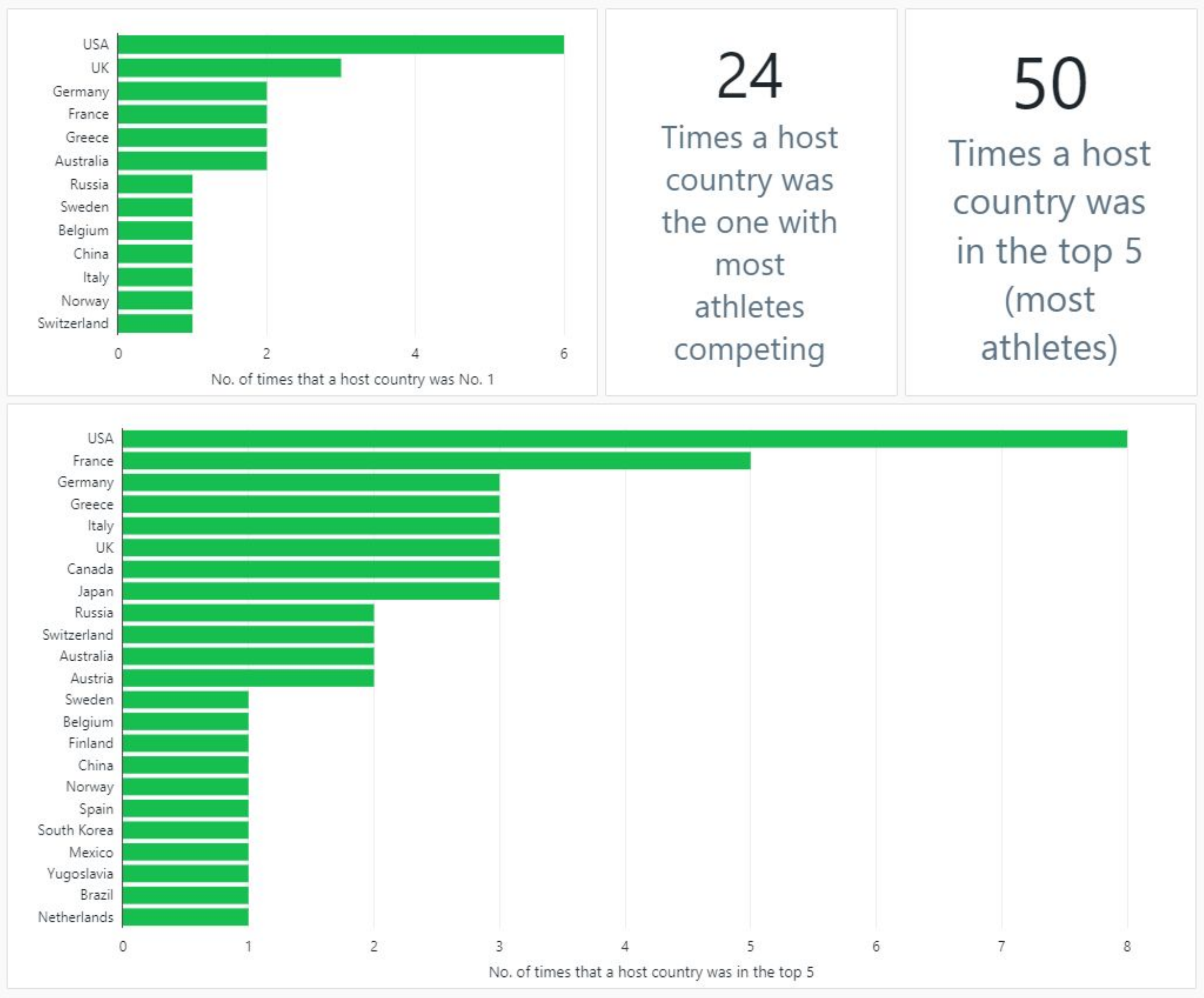
I looked at the **frequency** of the host country in the top 5; calculated athletes and medals **average**, **impact** of hosting (%), **athletes-medals correlation** and used ABBA tool to interpret results of **medal-athlete proportion** when hosting vs not hosting the games.



country	games	host	ranking	athletes
Greece	1896 Summer	Greece	1	102
Germany	1896 Summer	Greece	2	19
USA	1896 Summer	Greece	3	14
France	1896 Summer	Greece	4	12
UK	1896 Summer	Greece	5	10
France	1900 Summer	France	1	720
UK	1900 Summer	France	2	104
Germany	1900 Summer	France	3	76
USA	1900 Summer	France	4	74
Belgium	1900 Summer	France	5	64
USA	1904 Summer	USA	1	520
Canada	1904 Summer	USA	2	56
Germany	1904 Summer	USA	3	22
Greece	1904 Summer	USA	4	14
South Africa	1904 Summer	USA	5	8



country	games	host	ranking	medals
Greece	1896 Summer	Greece	1	48
Germany	1896 Summer	Greece	2	32
USA	1896 Summer	Greece	3	20
France	1896 Summer	Greece	4	11
UK	1896 Summer	Greece	5	9
France	1900 Summer	France	1	235
UK	1900 Summer	France	2	108
USA	1900 Summer	France	3	62
Germany	1900 Summer	France	4	45
Belgium	1900 Summer	France	5	43
USA	1904 Summer	USA	1	390
Canada	1904 Summer	USA	2	47
Germany	1904 Summer	USA	3	16
Cuba	1904 Summer	USA	4	5
Australia	1904 Summer	USA	5	4

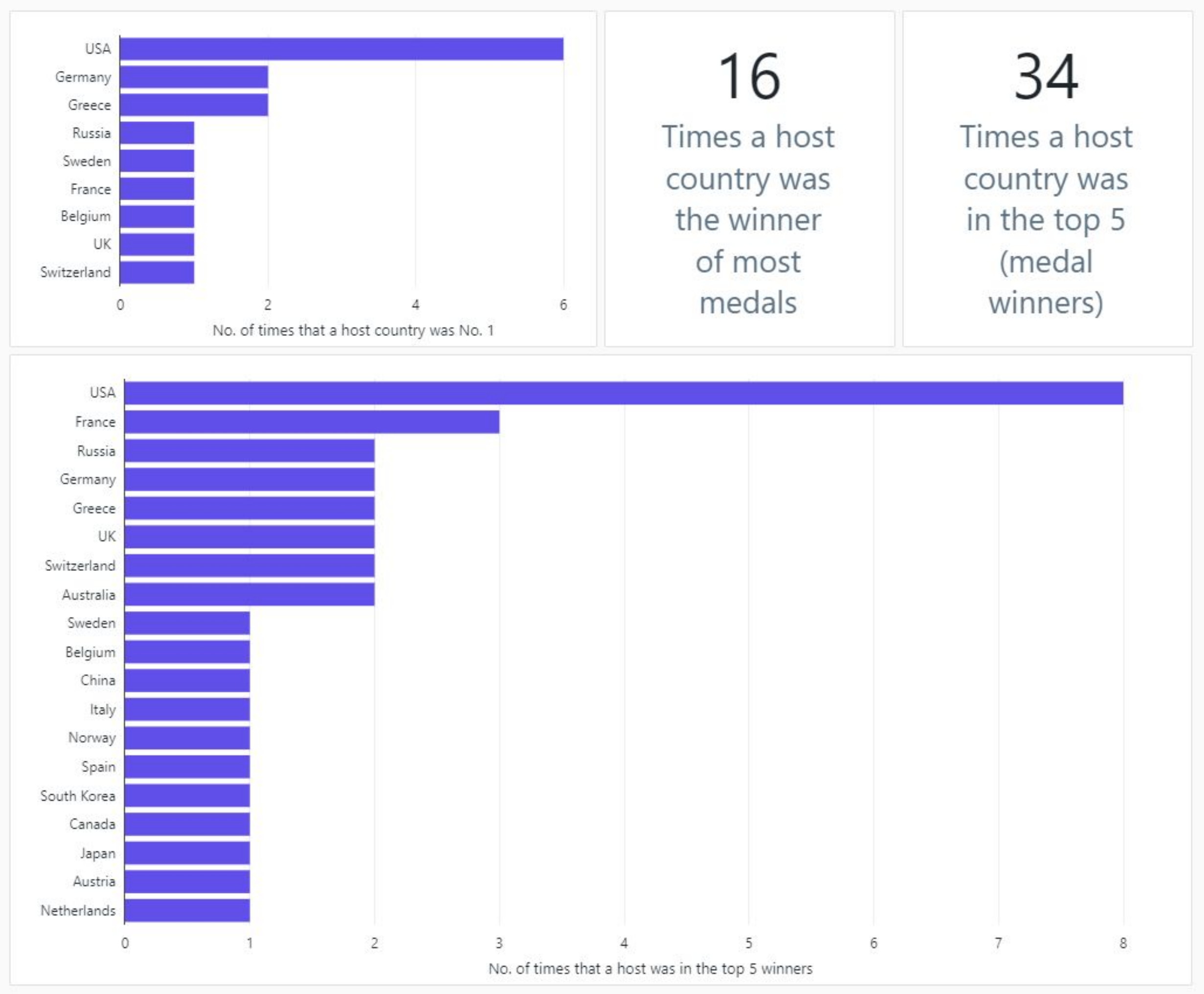


24

Times a host country was the one with most athletes competing

50

Times a host country was in the top 5 (most athletes)



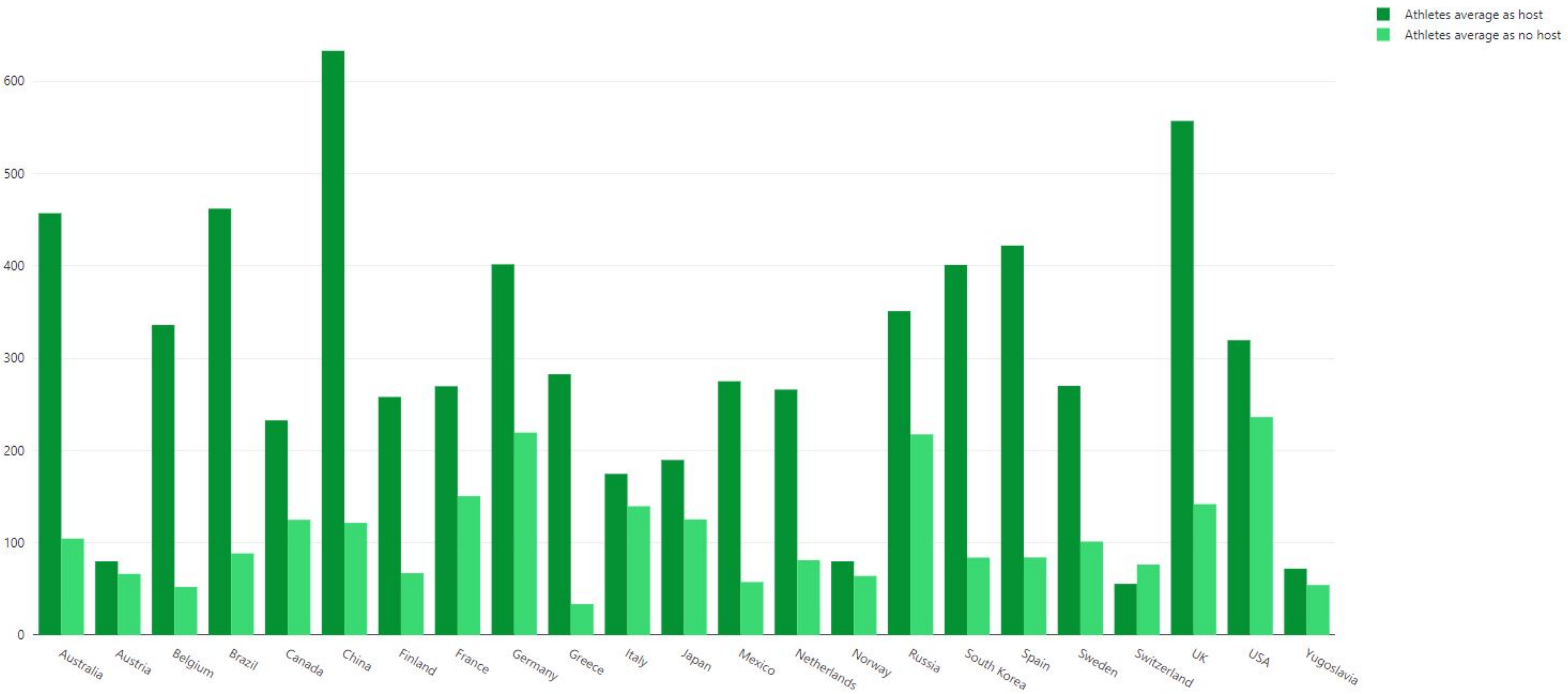
16

Times a host country was the winner of most medals

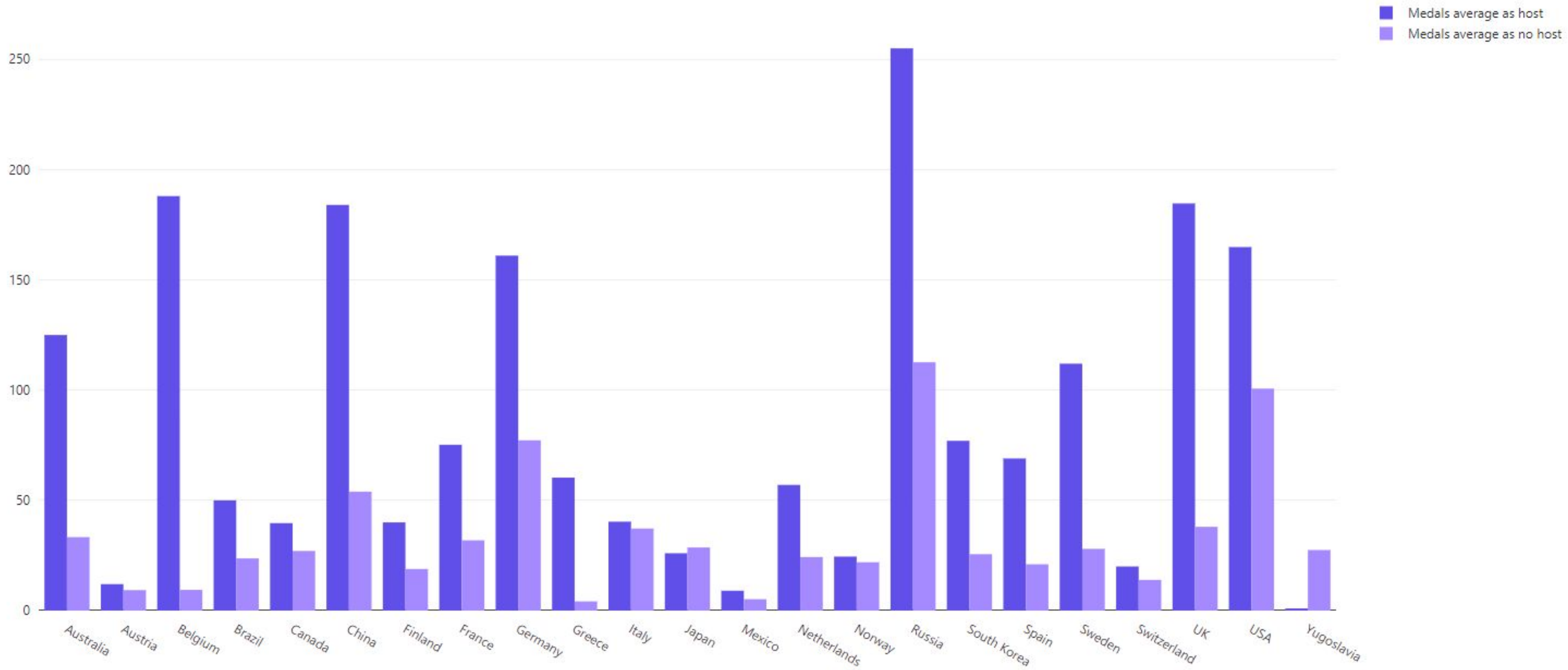
34

Times a host country was in the top 5 (medal winners)

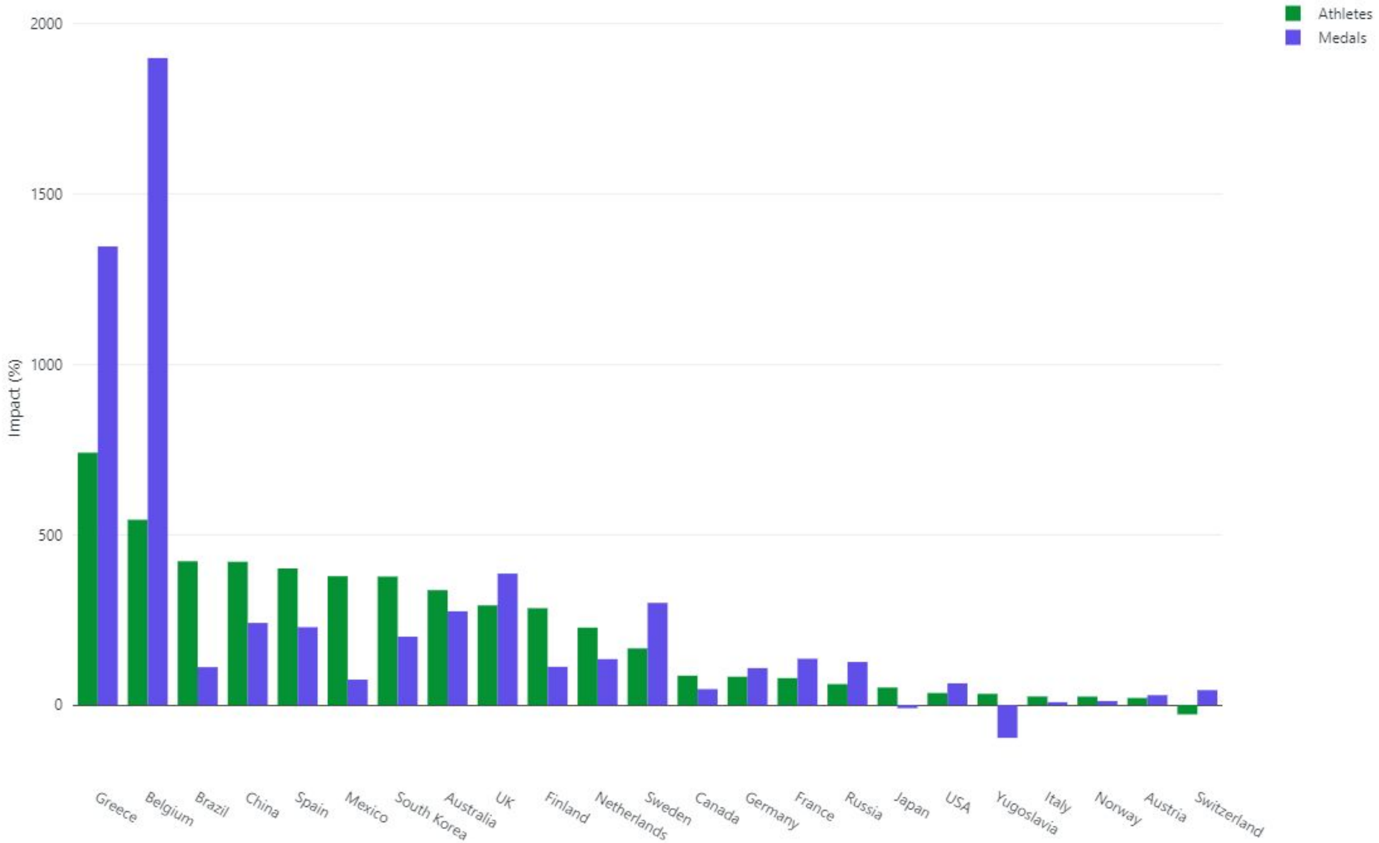
Higher number of athletes when a country was hosting the games (except for Switzerland)



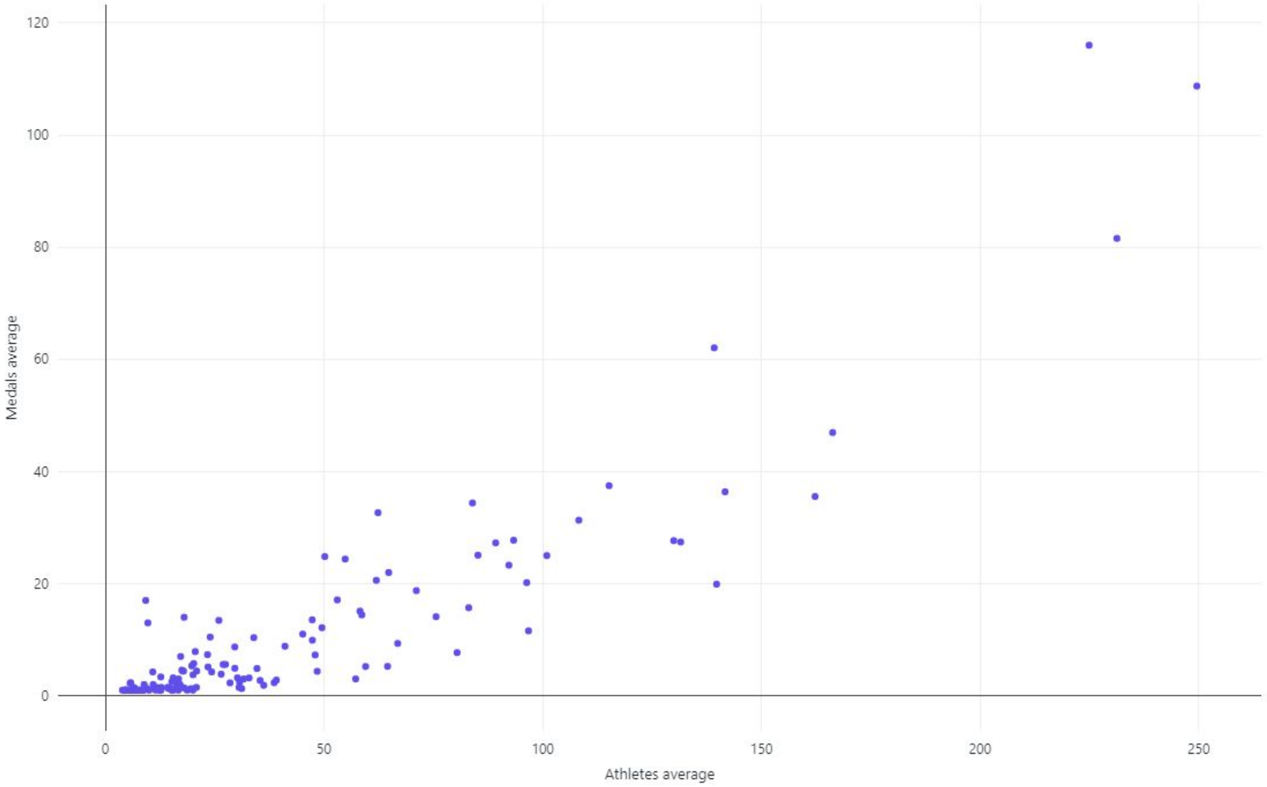
Higher number of medals when a country was hosting the games (except for Japan and Yugoslavia)



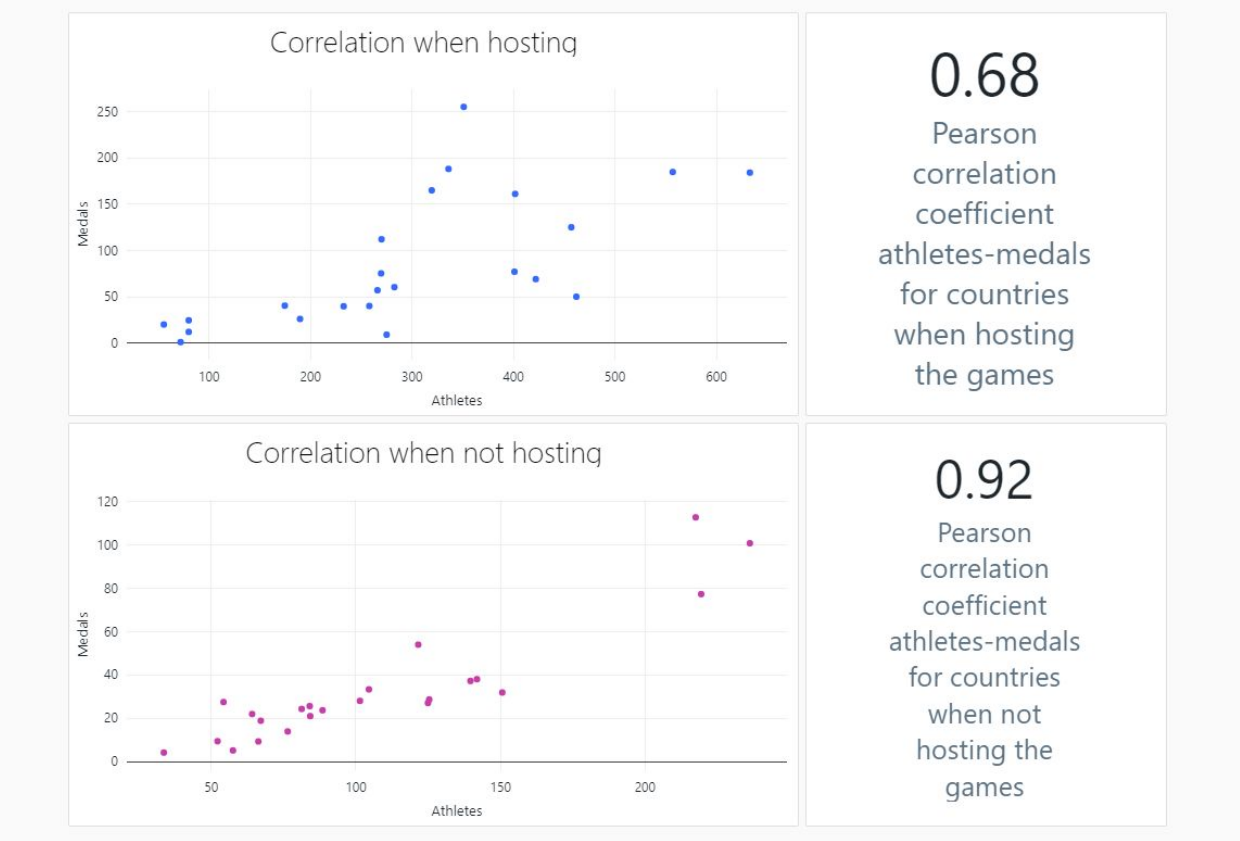
The greatest improvement in medals won while hosting was seen for Belgium (1900%) and Greece (1347%)



For reference, there is a clear correlation between athletes and medals won for all countries (**both host and non-host**)



Interestingly, this correlation still exists but is weaker when hosting the games



0.68

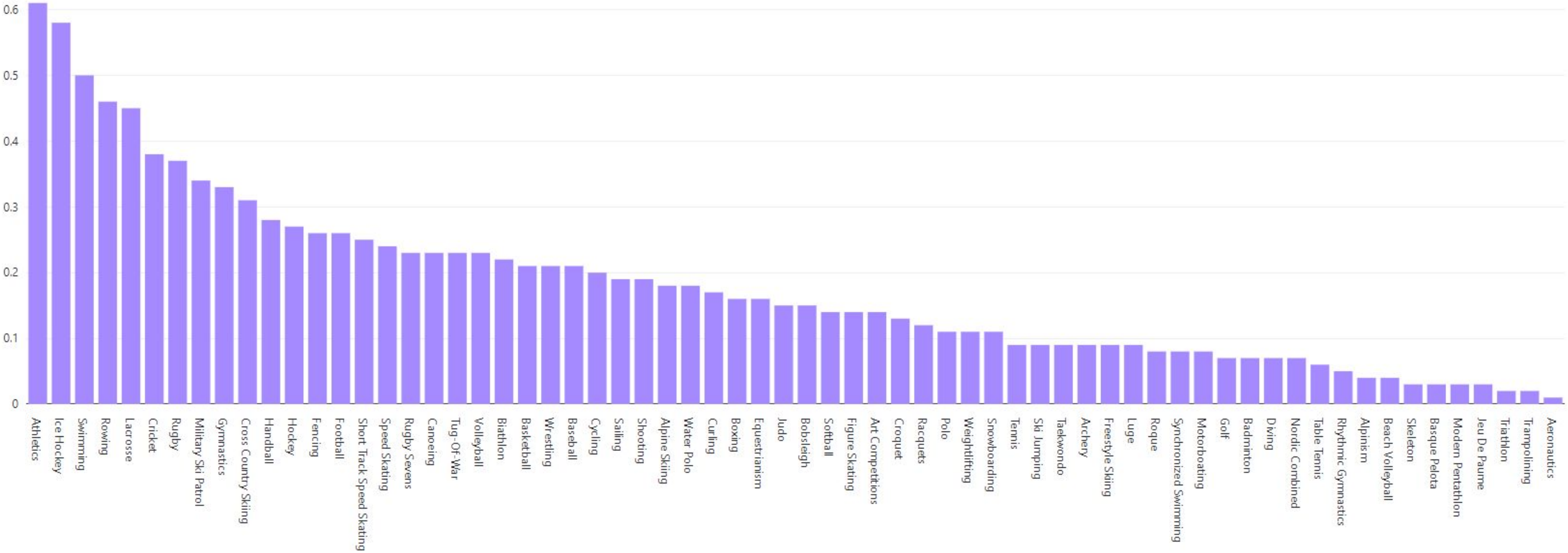
Pearson correlation coefficient athletes-medals for countries when hosting the games

0.92

Pearson correlation coefficient athletes-medals for countries when not hosting the games

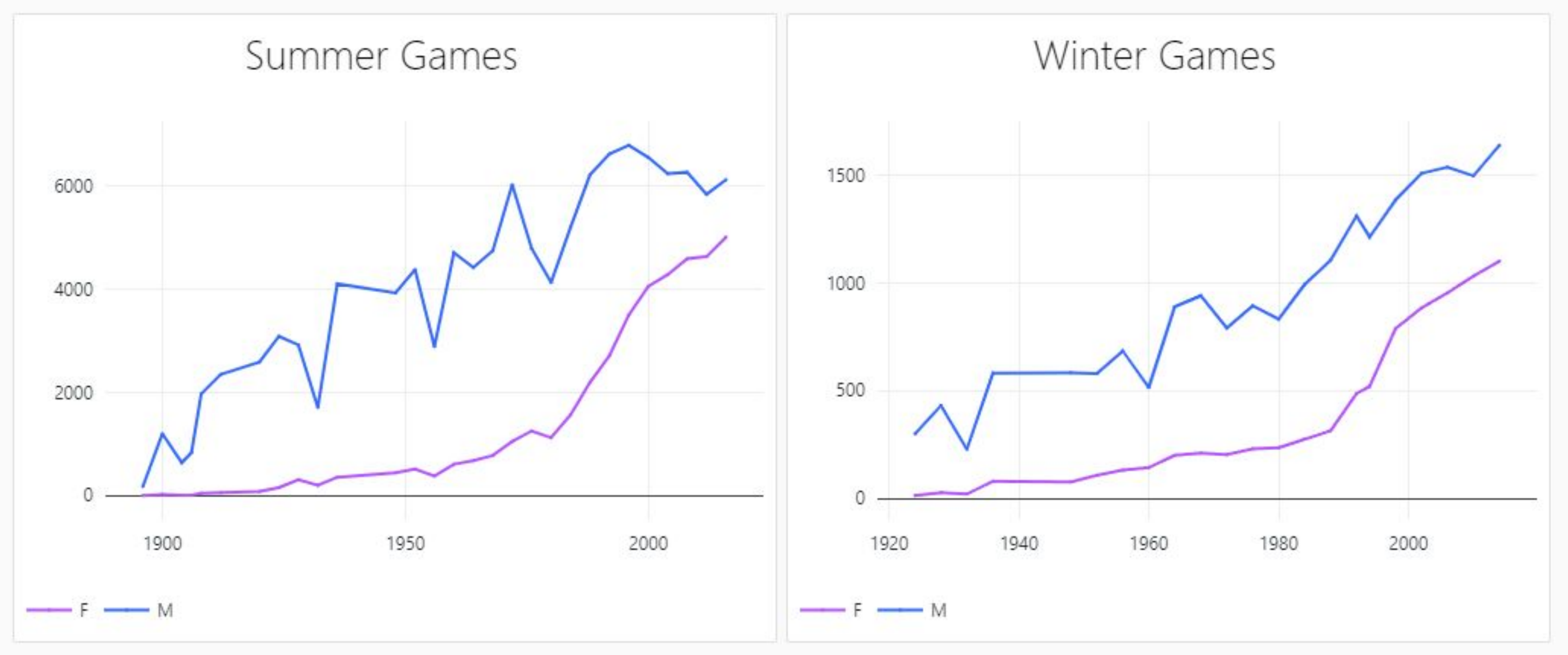
Analyzing medal-athlete ratio by sport

Athletics and **Swimming** are the top sport categories with more medals won, in proportion to the number of athletes

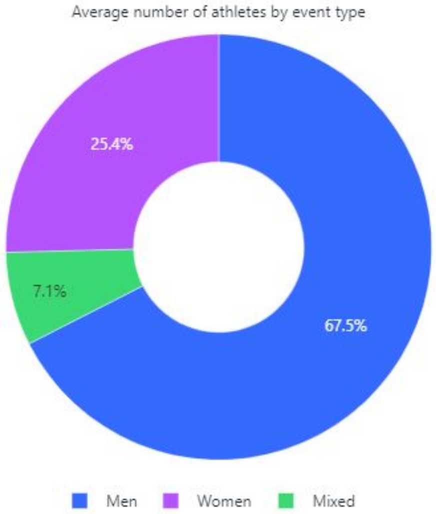


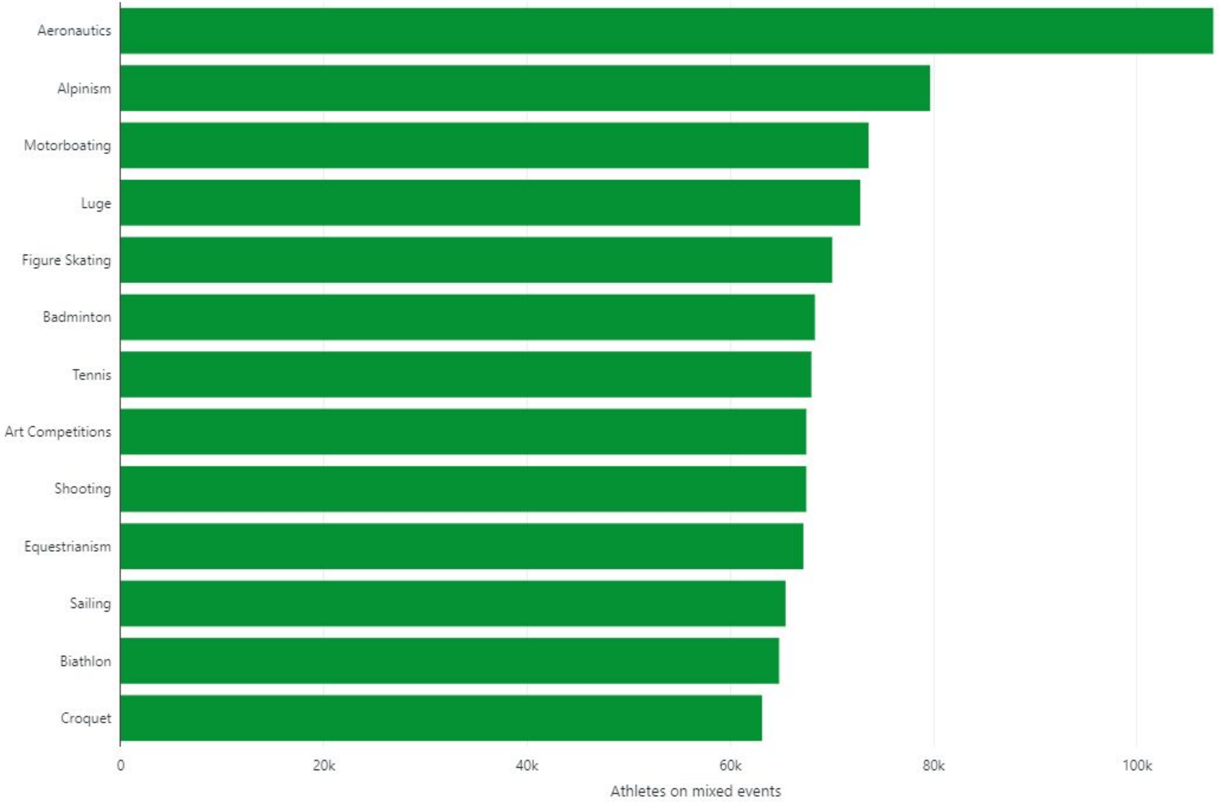
Analyzing athletes and medals by sex

There are more male than female athletes competing in the games, although this **disparity** has been decreasing over the years

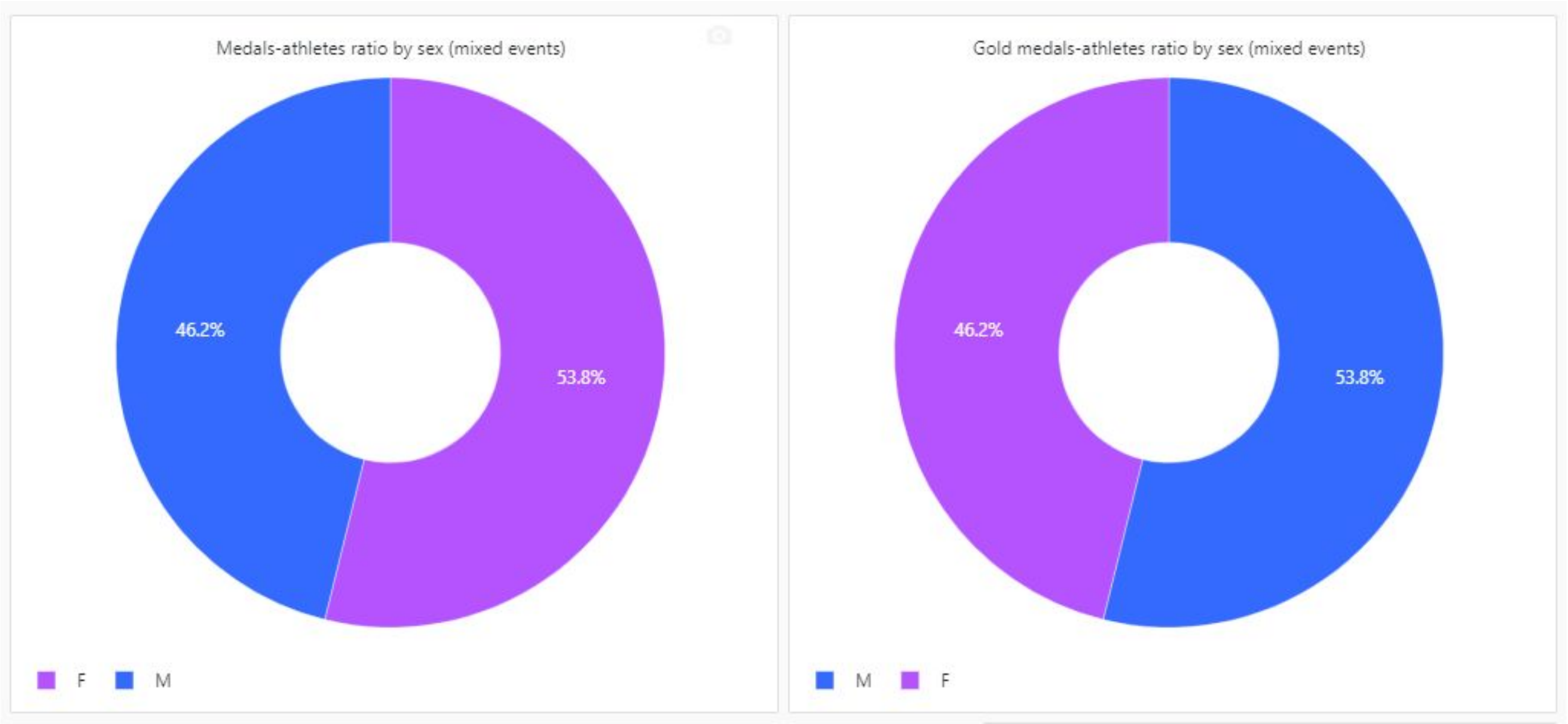


Most sport events are divided into categories by sex, but men and women can compete against each other on certain **mixed events**





In proportion to the number of athletes, women win more medals than men
But men win more gold medals

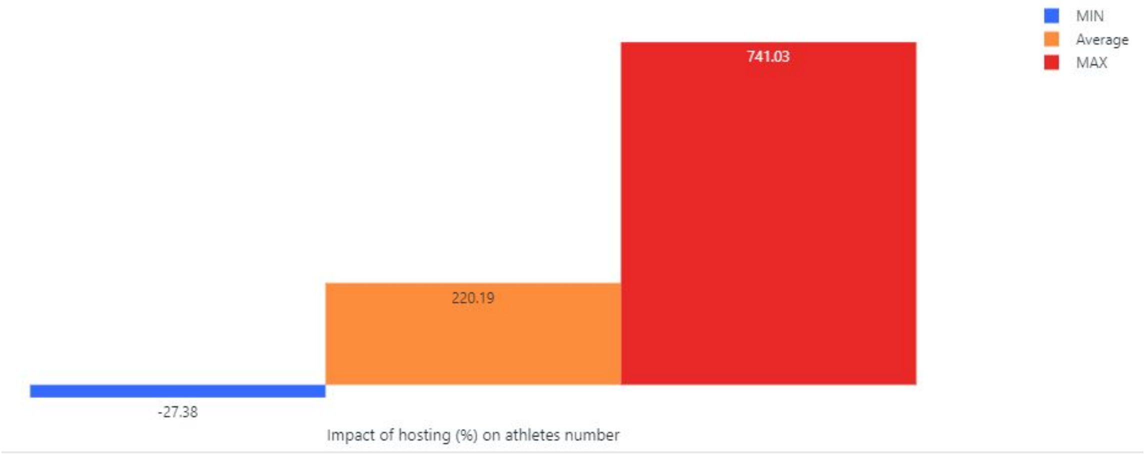


Hypotheses Results

Hosting the Olympics correlates to a higher number of athletes participating in the events

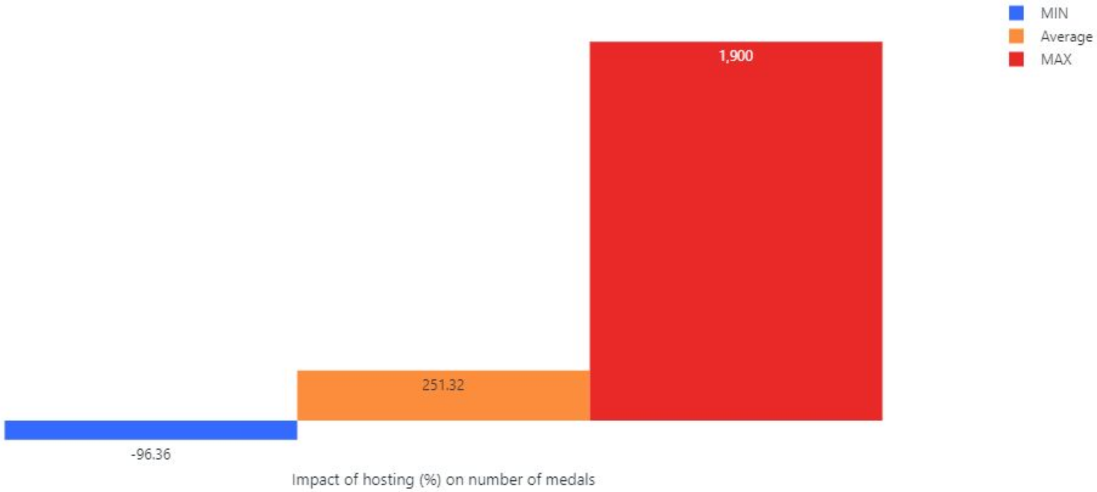
This was **true** for all countries but Switzerland (27% less athletes when hosting).

It would be interesting to investigate the root cause, and to further analyze this metric for the countries that are geographically closer to the host country. This may **benefit** the Olympics host's **neighbour countries**.



A higher number of athletes may contribute to winning more medals

There was a big positive **impact** on the number of **medals** won by the countries when hosting the games, except for Yugoslavia (96% less medals) and Japan (9% less medals).



- There was a **strong** positive **correlation** between athletes and medals when the country was **not hosting** the games (**0.92**), but a **weaker** positive **correlation** when the country was the **host** (**0.68**).

  This could indicate **diminishing returns** to the clear positive initial impact that an increased number of athletes seems to have on medal win rate, where eventually the **medal per athlete ratio** will decline.

- With that in mind, I also analyzed the **medal-athlete proportion** for these two scenarios and interpreted the results using the ABBA tool (statistical tool for analysis of binomial data): https://thumbtack.github.io/abba/demo/abba.html#No_host=773%2C2492&Host=1975%2C6845&abba%3AintervalConfidenceLevel=0.95&abba%3AuseMultipleTestCorrection=true

  *The result shows with **95.1% confidence** that there was a **7% decrease** in **medals per athlete** when the country was hosting the games, due to the excessive athlete count.*

- When **all the countries** (not only the hosts) are taken into account, there is a **strong** positive **correlation** between athletes and medals won (**0.90**).

  Therefore, investing in a higher number of athletes may **increase the winning probability** with diminishing returns.

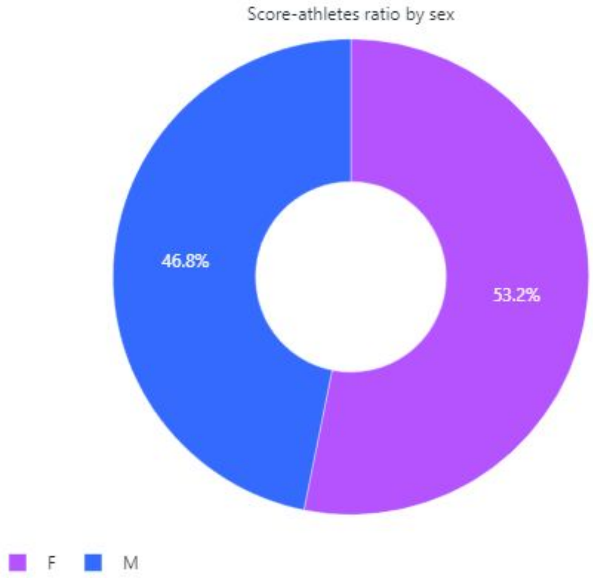
In proportion to the number of athletes, women win more medals than men

- On one hand, this hypothesis turned out to be **true**: 0.21 medals per **female** athlete and 0.18 medals per **male** athlete.
- On the other, the **gold**-medals ratio is slightly **higher** for **men** (0.07) than for women (0.06).
- To further analyze this hypothesis I created a score based upon a weighted point system, where Gold = 3 points, Silver = 2 points and Bronze = 1 point.

Based on this score in proportion to the number of athletes, women win more medals than men:

0.42 medals per **female** athlete and **0.37** medals per **male** athlete.

Having more female athletes competing on **mixed events** may be considered to slightly **increase** the **winning probability**.



Thank You!