Report Out: SportsStats

July 14, 2023

Daichi Azumi

Contents

Introduction

Background

- Selection of Client / Dataset

  Client: Supervisor in a data analysis company

  Dataset: SportsStats

- Situation

  As a junior data analyst, I assigned to analyze SportsStats data.
  My job is to analyze the dataset and find some insights, and then report out it to my supervisor.

Questions to Answer

1. Is there a relationship between gold medal winners and age?
   - If there is a relationship between gold medalists and their age, it may be helpful in the selection of national players and team composition.

2. Is there a relationship between gold medal winners and their countries?
   - If there is a relationship between gold medalists and their countries, the performance may be improved by investigating the features of the countries such as the climate and culture.

3. Which athletes have won the most medals?
   - Interviewing the unique training methods and mindsets of athletes who have won numerous medals may lead to improved performance.

Initial Hypotheses

1. Is there a relationship between gold medal winners and age?

(a) Gold medal winners are concentrated in the age range of 20 - 30 years old.

2. Is there a relationship between gold medal winners and their countries?

(b) The majority of gold medal winners are from the developed countries like the United States, Great Britain, Germany, and so on.

3. Which athletes have won the most medals?

(c) American athlete has won the most medals.

Data Analysis Approach

(a) Gold medal winners are concentrated in the age range of 20 - 30 years old.

- Group by age and count the number of medalists
- Visualize the counts

(b) The majority of gold medal winners are from the developed countries like the United States, Great Britain, Germany, and so on.

- Group by region and then visualize the relationship

(c) American athlete has won the most medals.

- Group by athlete ID and count the number of medals won, then sort by counts

Prep for Data Analysis

Import and Clean the Data

1. Open Jupyter Notebooks
2. Import the CSV files using Pandas function "read_csv"
3. I just imported two CSV files, "athlete_events.csv" and "noc_regions.csv".
   I wanted to join two tables for simplicity, so I decided to do JOIN operation in SQL.
4. I performed INNER JOIN using SQLite
5. To create a table containing only medalists, I specified the condition
   "Medal IS NOT NULL" in the WHERE clause and then created a new table named
   "olympic_medalists"

Initial Exploration

- Descriptive statistics (mean, mode, standard deviation, min, max, etc.)

```
olympic_medalists.describe(include='all')
✓ 0.0s
```

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	region	notes
count	39774.000000	39774	39774	39042.000000	31063.000000	30447.000000	39774	39774	39774	39774.000000	39774	39774	39774	39774	39774	39774	545
unique	NaN	28197	2	NaN	NaN	NaN	497	148	51	NaN	2	42	66	756	3	136	11
top	NaN	Michael Fred Phelps, II	M	NaN	NaN	NaN	United States	USA	2008 Summer	NaN	Summer	London	Athletics	Football Men's Football	Gold	USA	Yugoslavia
freq	NaN	28	28528	NaN	NaN	NaN	5219	5637	2045	NaN	34079	3620	3969	1269	13371	5637	390
mean	69404.520139	NaN	NaN	25.925132	177.557898	73.774274	NaN	NaN	NaN	1973.936743	NaN	NaN	NaN	NaN	NaN	NaN	NaN
std	38849.700624	NaN	NaN	5.914471	10.892172	15.016399	NaN	NaN	NaN	33.822507	NaN	NaN	NaN	NaN	NaN	NaN	NaN
min	4.000000	NaN	NaN	10.000000	136.000000	28.000000	NaN	NaN	NaN	1896.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN
25%	36496.500000	NaN	NaN	22.000000	170.000000	63.000000	NaN	NaN	NaN	1952.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN
50%	68985.000000	NaN	NaN	25.000000	178.000000	73.000000	NaN	NaN	NaN	1984.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN
75%	103459.750000	NaN	NaN	29.000000	185.000000	83.000000	NaN	NaN	NaN	2002.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN
max	135563.000000	NaN	NaN	73.000000	223.000000	182.000000	NaN	NaN	NaN	2016.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN

- The number of medals by medal colors

```
%%sql

SELECT
    Medal,
    COUNT(*) AS count
FROM
    olympic_medalists
GROUP BY
    Medal
ORDER BY
    count DESC
✓ 0.0s
```

Medal	count
Gold	13371
Bronze	13291
Silver	13112

- The number of medals by events

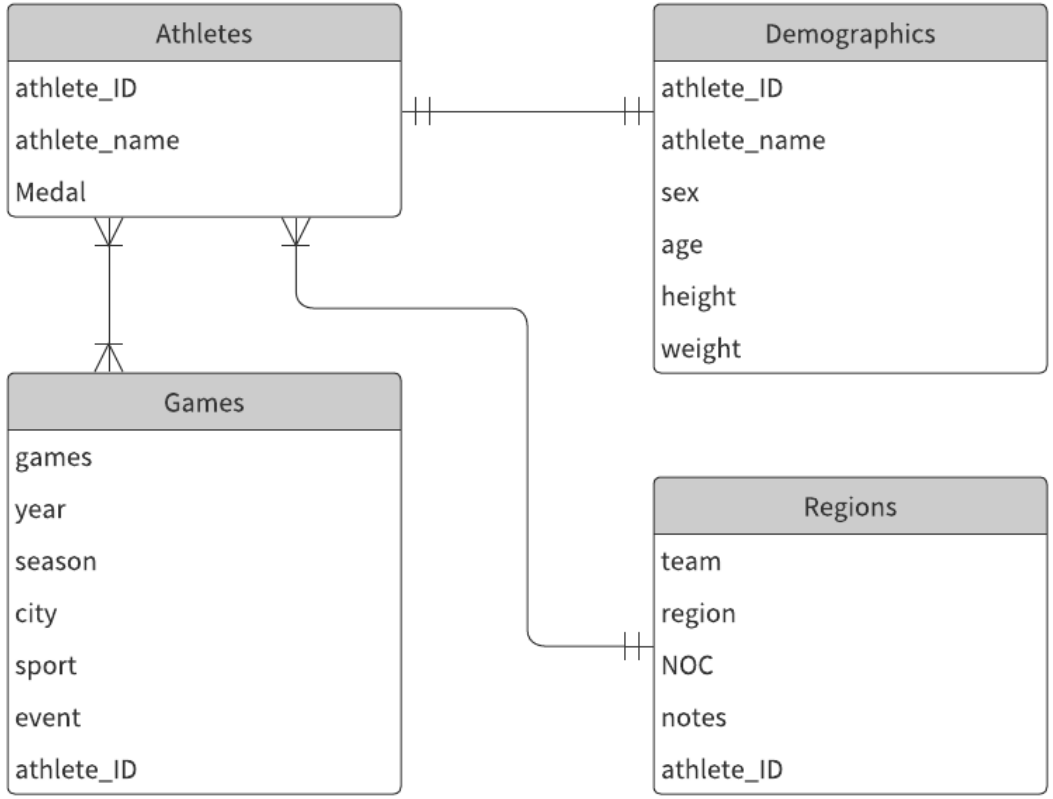
```
%%sql

SELECT
    Event,
    COUNT(*) AS count
FROM
    olympic_medalists
GROUP BY
    Event
ORDER BY
    count DESC
LIMIT
    10
✓ 0.0s
```

Event	count
Football Men's Football	1269
Ice Hockey Men's Ice Hockey	1230
Hockey Men's Hockey	1050
Water Polo Men's Water Polo	866
Rowing Men's Coxed Eights	730
Gymnastics Men's Team All-Around	713
Basketball Men's Basketball	687
Handball Men's Handball	588
Volleyball Men's Volleyball	495
Hockey Women's Hockey	478

Initial Findings

- From the "Age" column, we can see that the minimum value is 10. In other words, a 10-year-old child won a medal.
- The mean age of medalists is 25.925132 and the standard deviation is 5.914471.
- From 1896 to 2016, the total number of gold, silver, and bronze medals are 13371, 13112, and 13291, respectively.
- From the "Team" column, it can be said that 497 teams have participated and the United States team won the most medals.

Entity Relationship Diagram

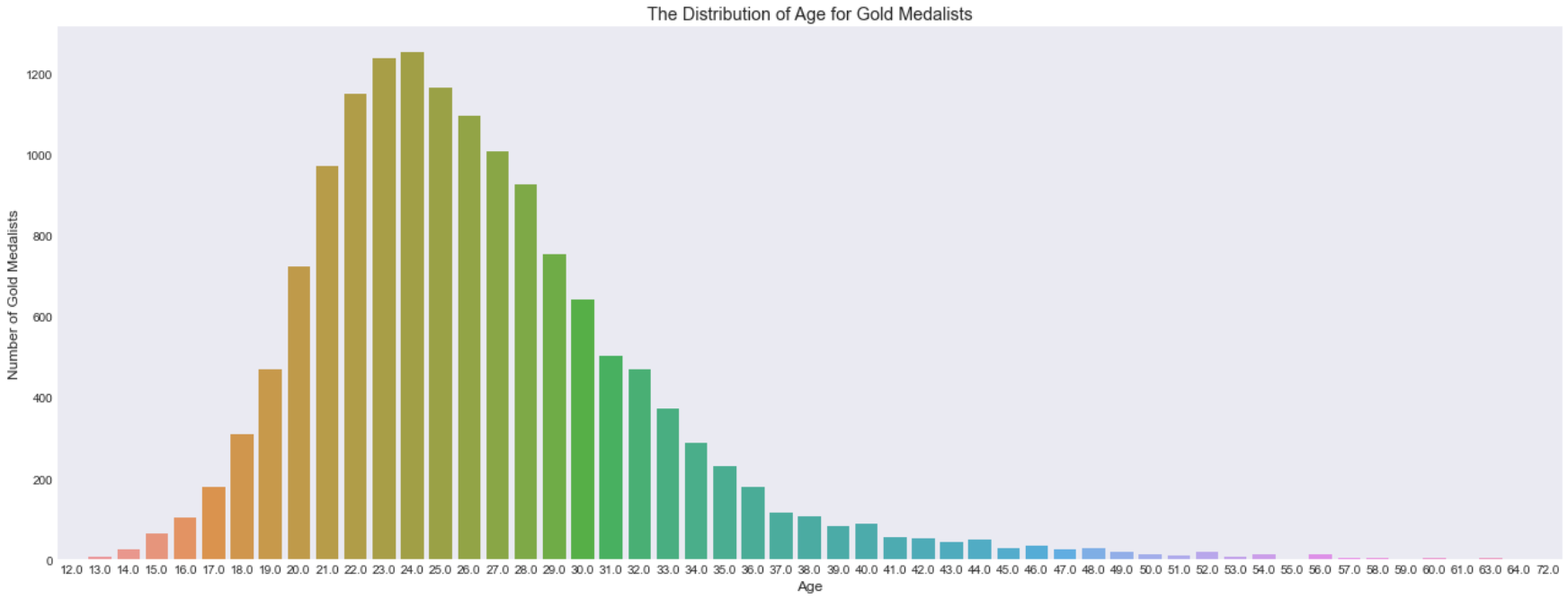


Data Analysis

Verification of Hypotheses (1)

(a) Gold medal winners are concentrated in the age range of 20 - 30 years old.

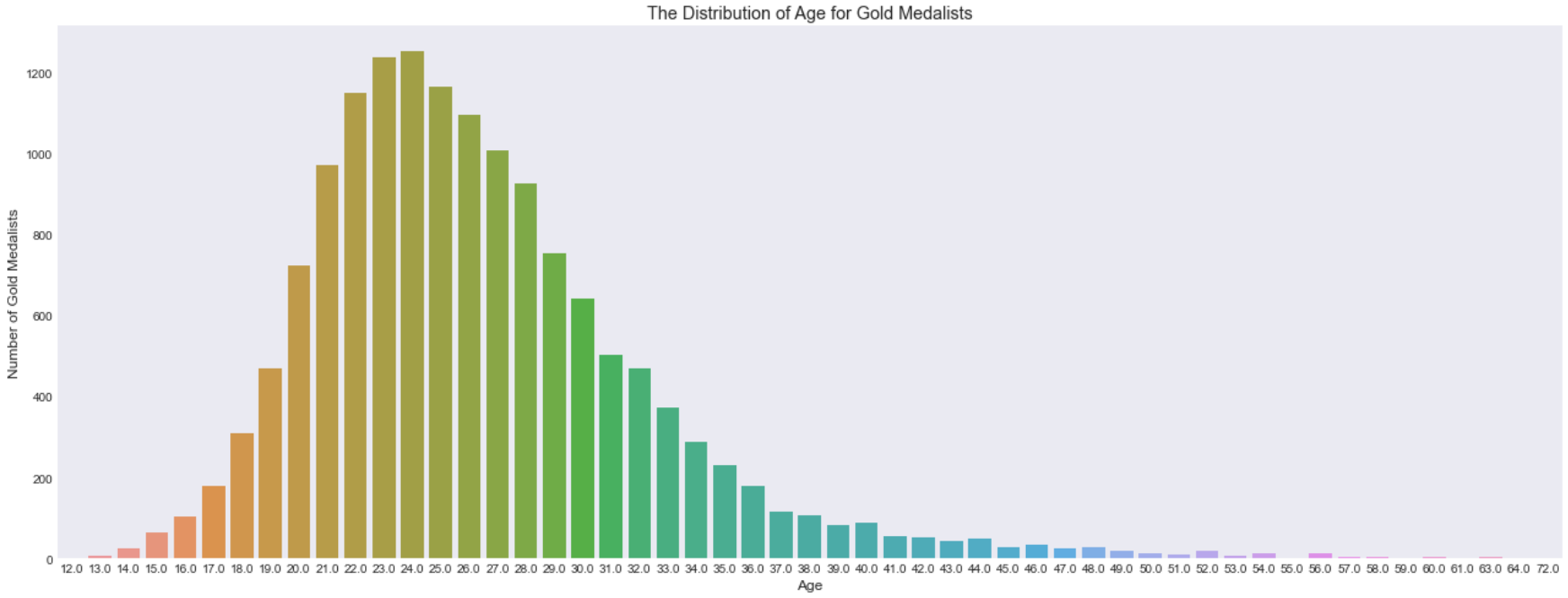
Proved. The ages of the gold medalists are mostly concentrated in the 20 to 30 age range.



Verification of Hypotheses (2)

(a) Gold medal winners are concentrated in the age range of 20 - 30 years old.

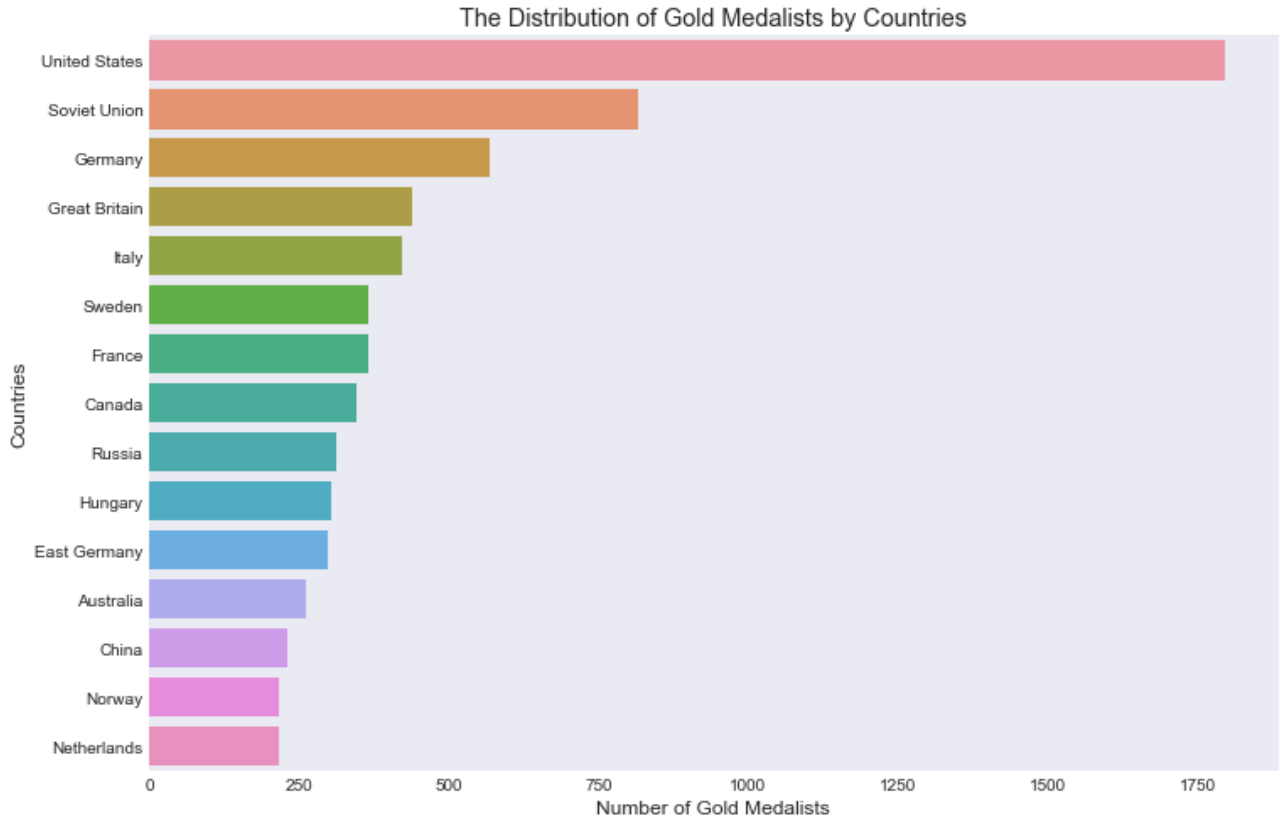
Then, how about the age distribution of silver and bronze medalists? Does it change little or does the peak shift to younger ages? To make sure that, I will perform additional analysis; What is the age distribution of silver and bronze medal winners?



Verification of Hypotheses (3)

(b) The majority of gold medal winners are from the developed countries like the United States, Great Britain, Germany, and so on.

Proved. The countries of gold medalists are, in descending order of the number of the gold medalists, the United States, the Soviet Union, Germany, Great Britain, Italy, and so on.



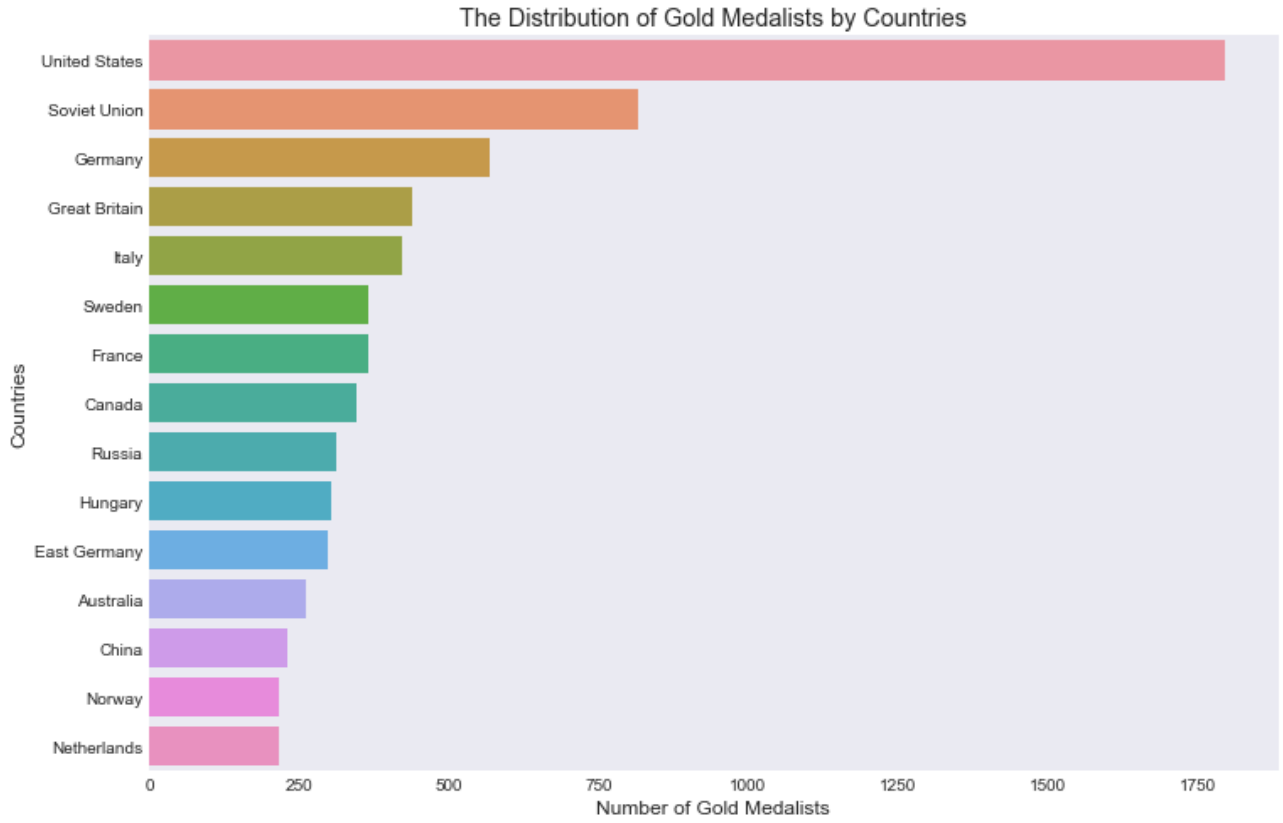
Verification of Hypotheses (4)

(b) The majority of gold medal winners are from the developed countries like the United States, Great Britain, Germany, and so on.

In addition to this fact, I would like to confirm that the countries of silver and bronze medalists have similar to the countries of gold medalists.

Thus, I will investigate the following question;

What is the distribution of the countries to which the silver and bronze medal winners belong?



Verification of Hypotheses (5)

(c) American athlete has won the most medals.

Proved. American athlete has won 28 medals.

However, when I looked at the medal count order, the percentage of American athletes was not so large in top 10.

Thus, I will make the following analysis;

What countries do the athletes who won more than one medals belong to?
What percentage of athletes belong to the United States among them?

athlete_ID	athlete_name	team	count_medals
94406	Michael Fred Phelps, II	United States	28
67046	Larysa Semenivna Latynina (Diriy-)	Soviet Union	18
4198	Nikolay Yefimovich Andrianov	Soviet Union	15
11951	Ole Einar Bjrndalen	Norway	13
74420	Edoardo Mangiarotti	Italy	13
89187	Takashi Ono	Japan	13
109161	Borys Anfiyanovych Shakhlin	Soviet Union	13
23426	Natalie Anne Coughlin (-Hall)	United States	12
35550	Birgit Fischer-Schmidt	East Germany	12
35550	Birgit Fischer-Schmidt	Germany	12

Additional Questions

1. Is there a relationship between the year of the Olympics and the number of medalists?

2. Is there a relationship between the year of the Olympics and the number of events?

3. What is the age distribution of silver and bronze medal winners?

4. What is the distribution of the countries to which the silver and bronze medal winners belong?

5. What country do the athletes who won more than one medals belong to? What percentage of athletes belong to the United States among them?

New Metric

I created some metrics for further analysis.

- number_of_medals

  This metric stores the number of medals an athlete won. I created this to extract the medalists with more than a given number of medals.

- ratio_medalists_country

  This metric represents the ratio of medalists from each country to the total number of medalists with more than one medal. I created this to visualize how the countries of medalists with more than one medal are distributed.

- ratio_medalists_participants

  This metric represents the ratio of medalists to total participants. I created this to visualize how the difficulty to be medalists changed through years.

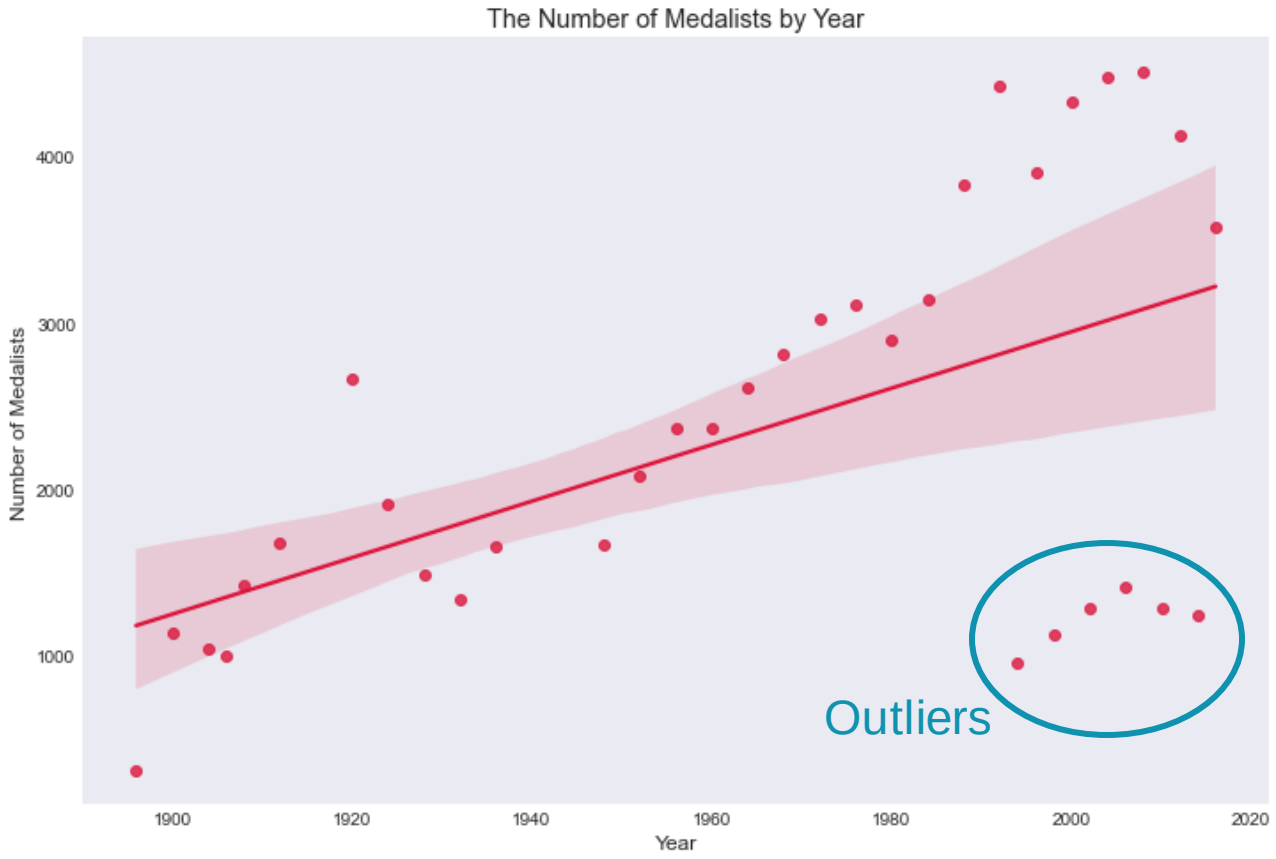
Deeper Analysis (1)

1. Is there a relationship between the year of the Olympics and the number of medalists?

As you can see, there is a correlation between year and the number of medalists.

However, there are outlier data from 1994.

However, some outliers are observed after 1994.



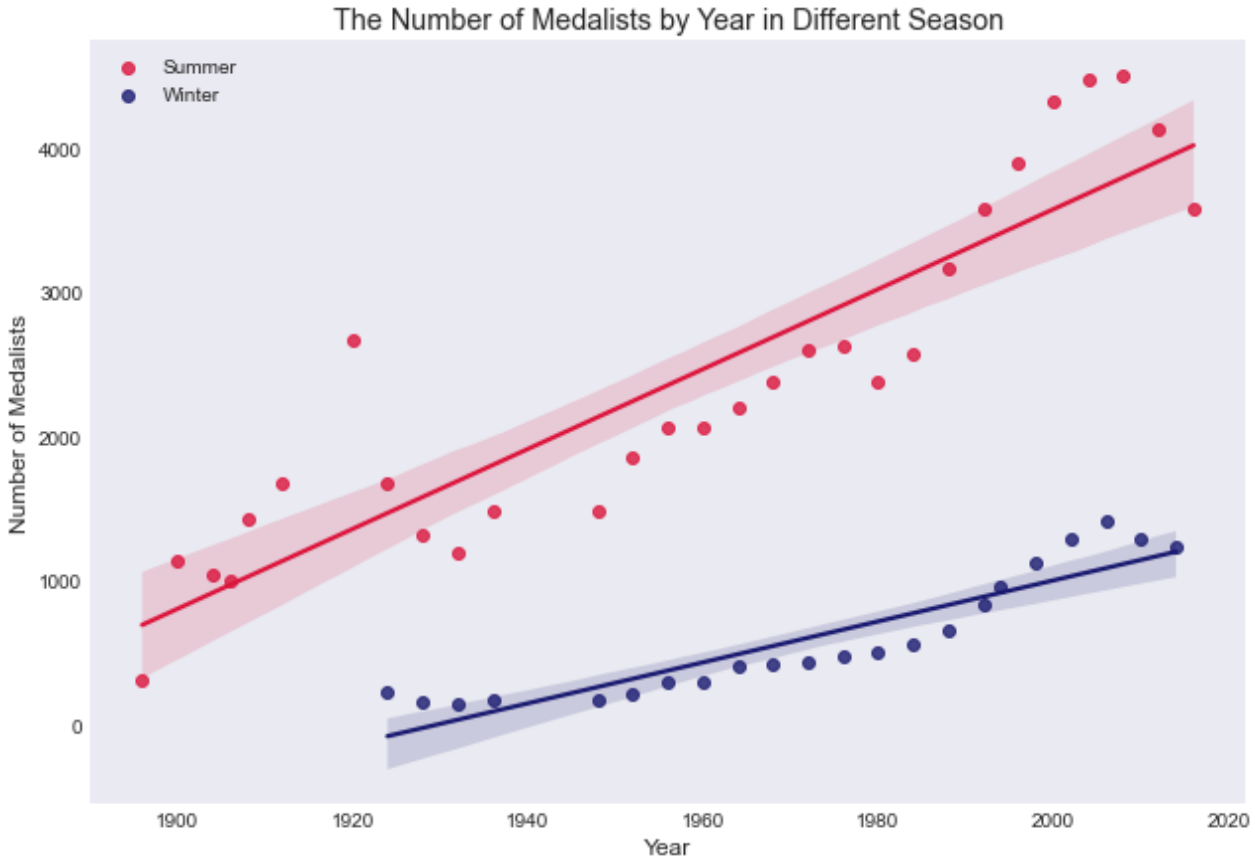
Deeper Analysis (2)

1. Is there a relationship between the year of the Olympics and the number of medalists?

To confirm that the outliers were due to seasonality, I separated the data by season.

The graph clearly shows that the outliers were due to seasonal differences.

The graph also shows that there is a correlation between the year and the number of medalists in both summer and winter.

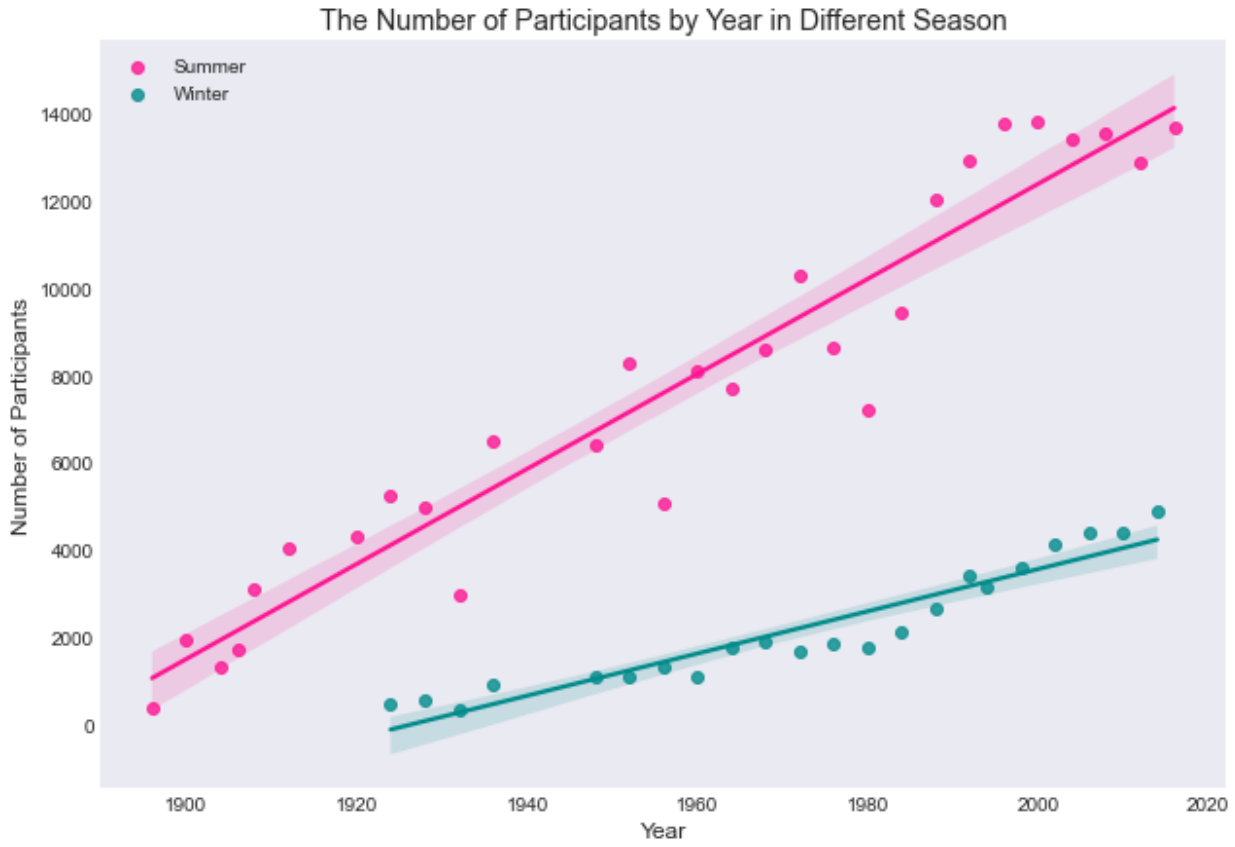


Deeper Analysis (3)

1. Is there a relationship between the year of the Olympics and the number of medalists?

The next step is to verify the relationship between the year and the number of Olympic participants.

The graph shows that there is a correlation between the year and the number of participants.



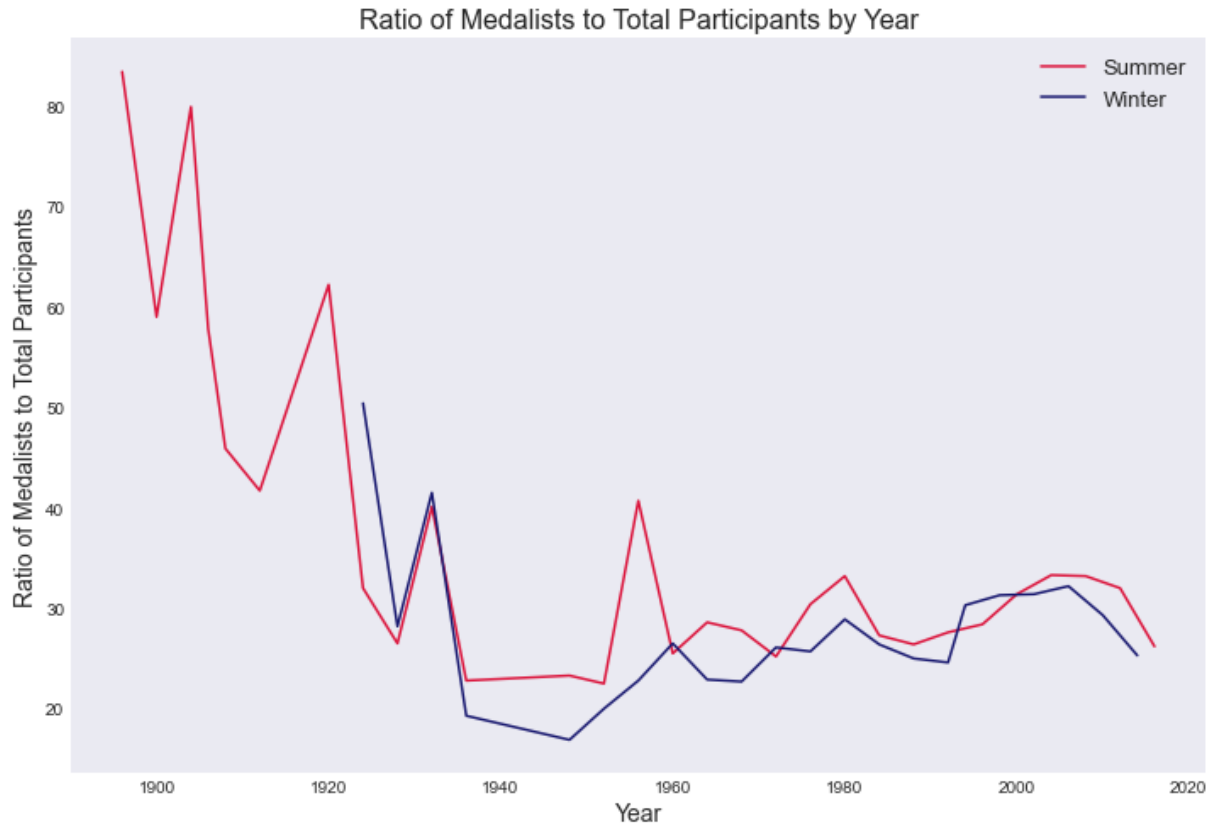
Deeper Analysis (4)

1. Is there a relationship between the year of the Olympics and the number of medalists?

By comparing the number of participants with the number of medalists, we can see how difficult to win medals by year.

The graph shows the ratio of medalists to total participants in both summer and winter by year.

According to the graph, the percentage of medalists continued to decline until around 1940, and since then has fluctuated between 20% and 30%.



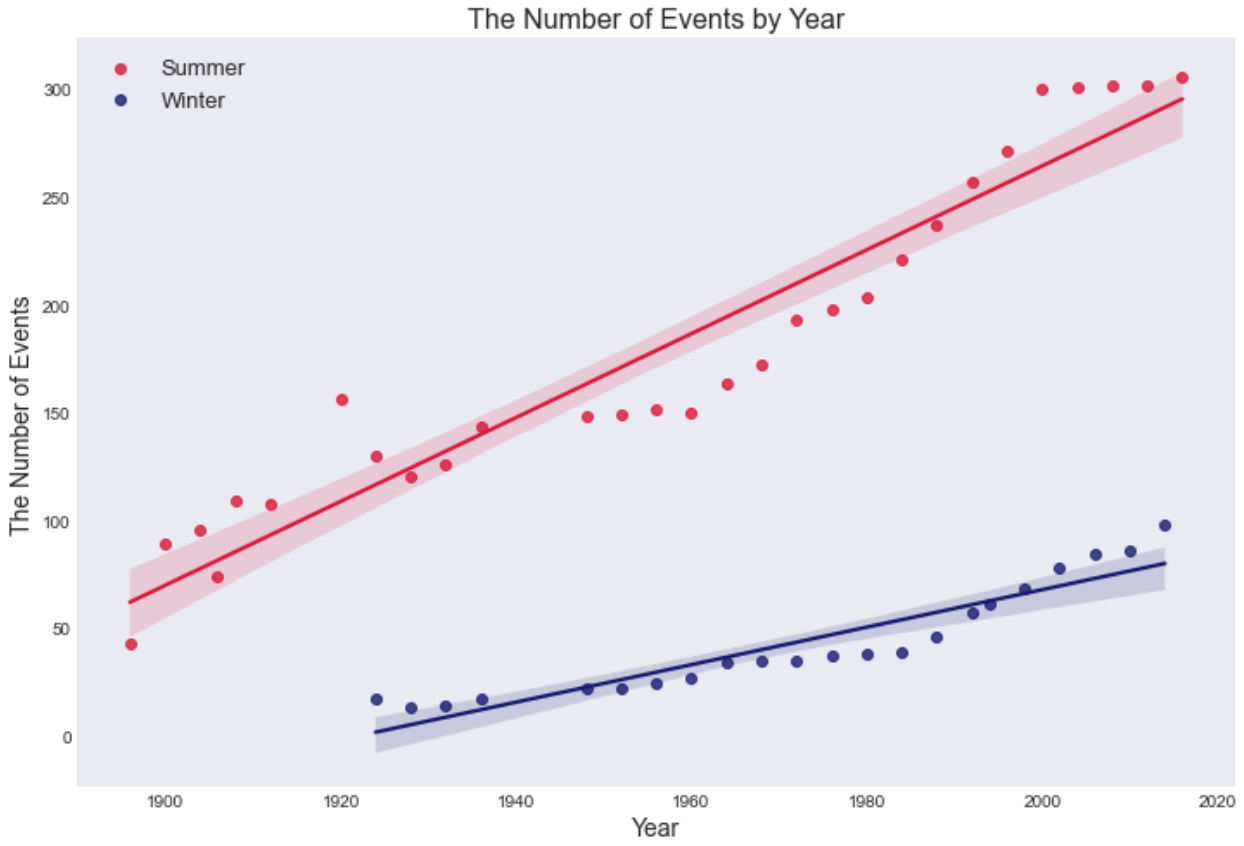
Deeper Analysis (5)

2. Is there a relationship between the year of the Olympics and the number of events?

The increase of medalists each year suggests that the number of events being held is increasing.

The graph confirms this inference.

And, of course, we can see that the rate of increase in medalists and the slope of the graph are almost identical.



Deeper Analysis (6)

3. What is the age distribution of silver and bronze medal winners?

In a previous verification of the hypothesis, I visualized the distribution of the age of the gold medalists. Following this, I visualized the age distribution of the silver and bronze medalists in order to compare the age distribution by medal color.

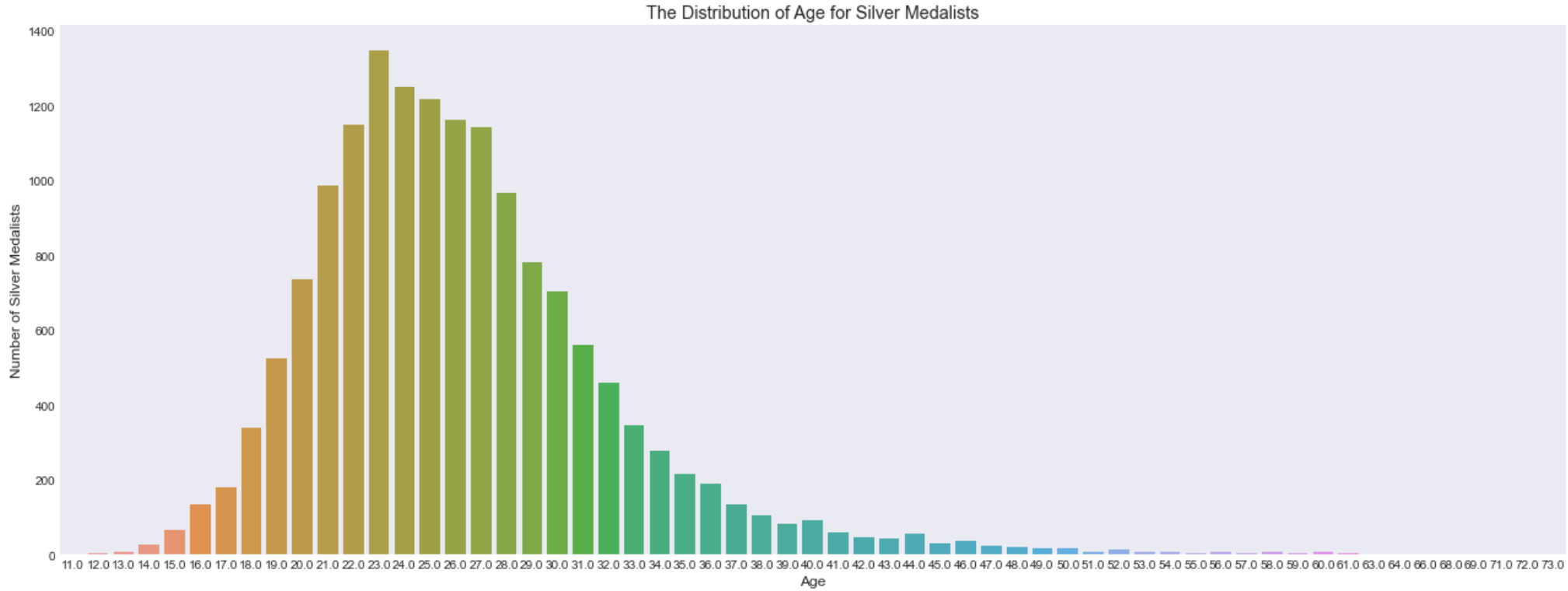
My expectation was that as the medal ranks increased, so would the age peak.

However, the results showed that the distribution of age varied quite little with medal color, although there was a slight difference.

Deeper Analysis (7)

3. What is the age distribution of silver and bronze medal winners?

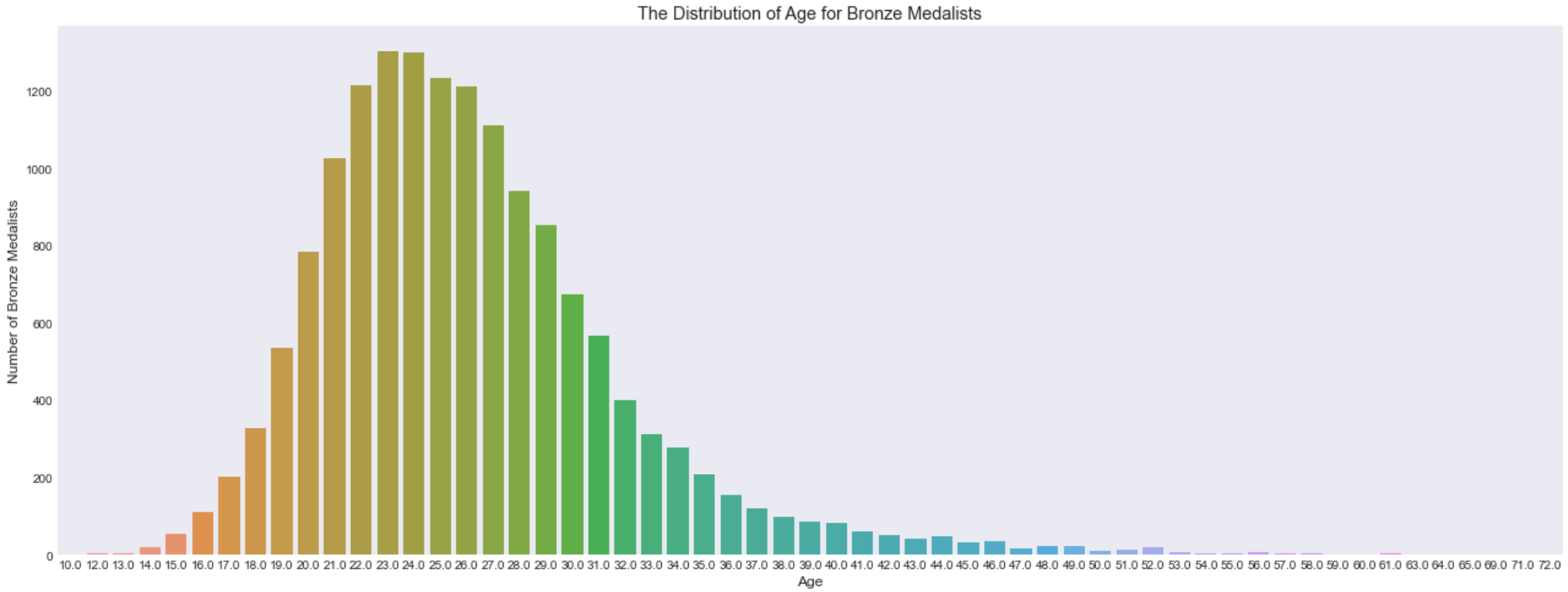
The graph shows the distribution of age for silver medalists.



Deeper Analysis (8)

3. What is the age distribution of silver and bronze medal winners?

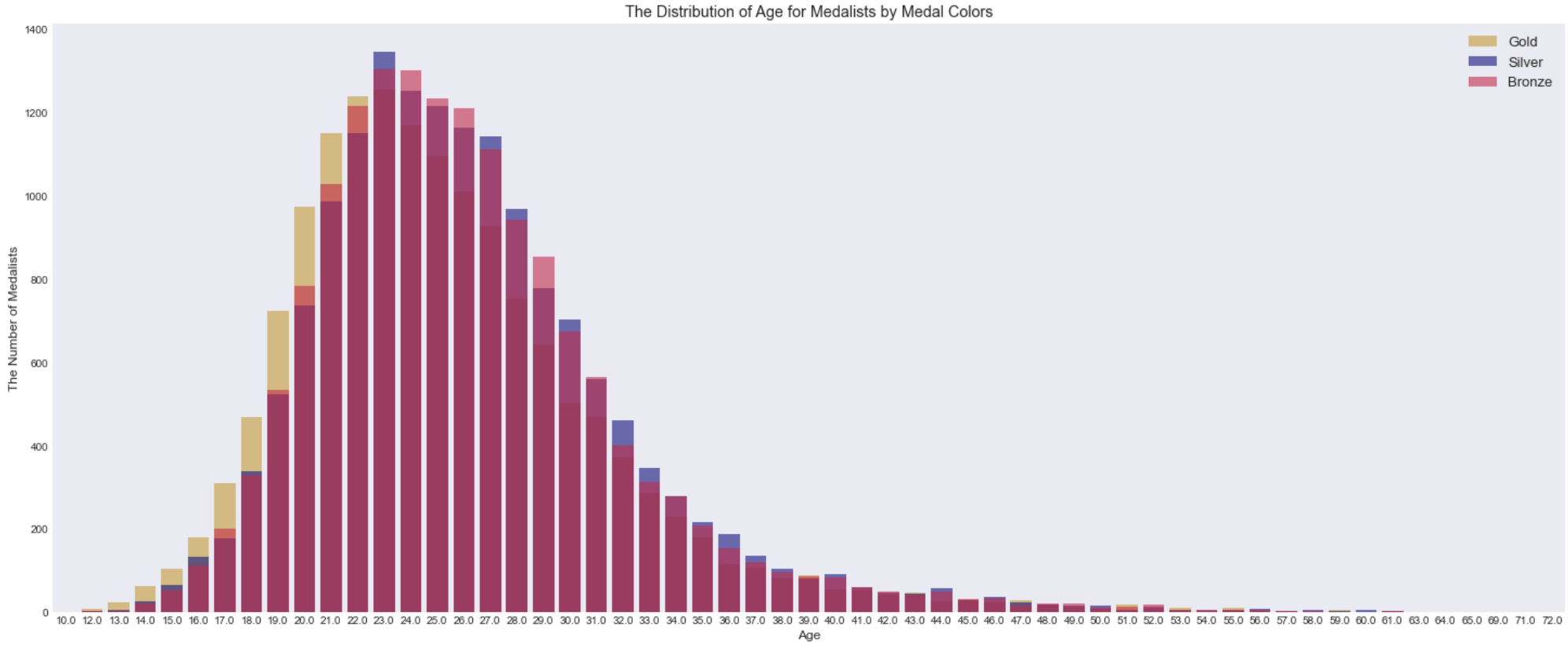
The graph shows the distribution of age for bronze medalists.



Deeper Analysis (9)

3. What is the age distribution of silver and bronze medal winners?

The graph shows the distribution of age for medalists by medal colors.



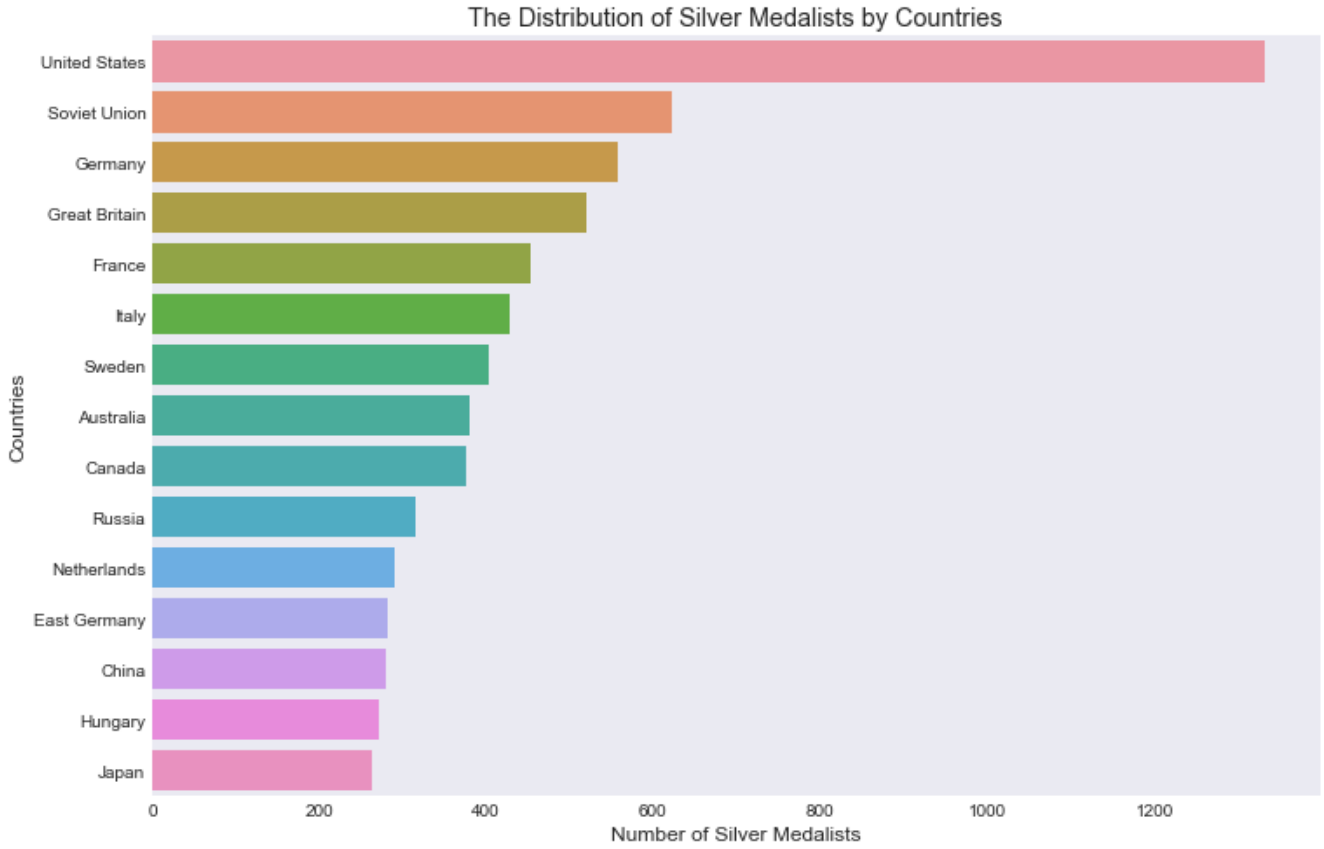
Deeper Analysis (10)

4. What is the distribution of the countries to which the silver and bronze medal winners belong?

The graph shows the distribution of silver medalists.

As I expected, the countries with most silver medalists are similar to the countries with most gold medalists.

The number of U.S. medalists is similarly high, but the percentage of medalists from other countries is increased overall.



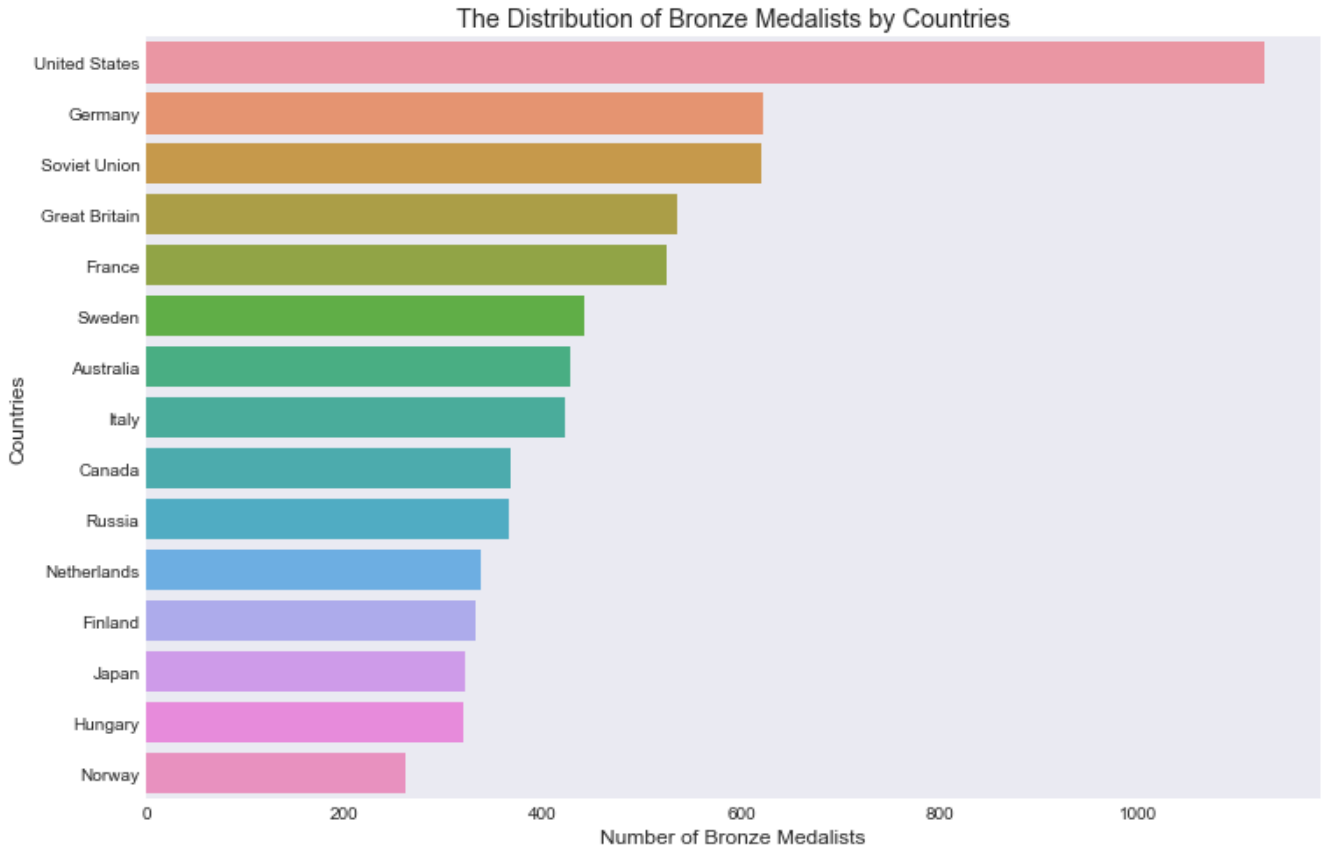
Deeper Analysis (11)

4. What is the distribution of the countries to which the silver and bronze medal winners belong?

The graph shows the distribution of bronze medalists.

The country lineup doesn't change that much.

The percentage of countries other than the U.S. is increased even more than in the case of silver medalists.



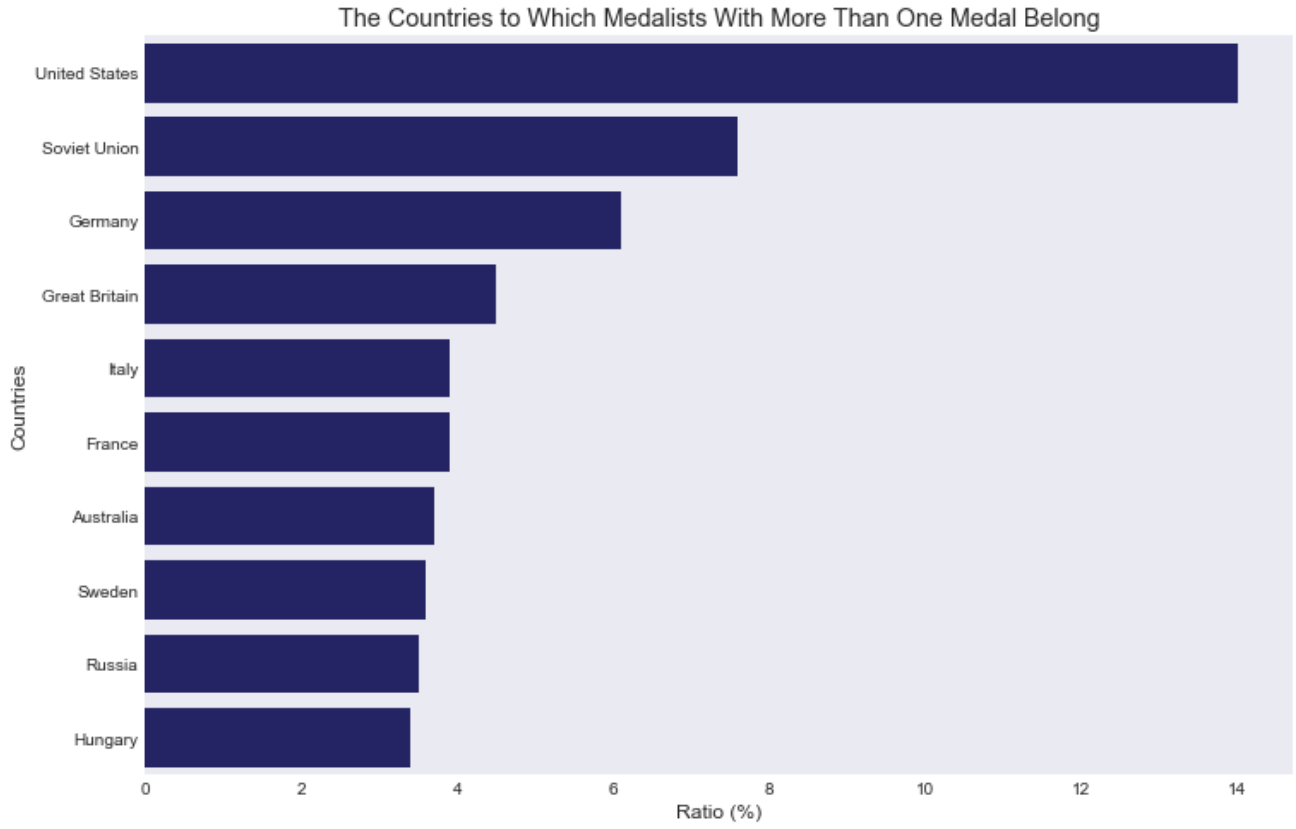
Deeper Analysis (12)

5. What country do the athletes who won more than one medals belong to? What percentage of athletes belong to the United States among them?

The graph shows the countries to which medalists with more than one medal belong.

The countries with multiple medal winners are, in descending order, the United States, the Soviet Union, Germany, Great Britain, and Italy.

The percentage of U.S. medalists is approximately 14%.



Conclusion and Proposal

Final Findings (1)

- The ages of the medalists are mostly concentrated in the range of 20 - 30 years old. Although there was a slight difference, the distribution of age varied quite little with medal colors.
- The majority of the medal winners were from developed countries such as the United States, Germany, and Great Britain. Most of the medalists were from the United States, but the percentage of medalists from other countries increased from gold to bronze medals.
- Similarly, the medalists with more than one medal tend to be from developed countries. The ratio of the U.S. medalists to the total number of these medalists was about 14%.

Final Findings (2)

- There is a positive correlation between the year of the Olympics and the number of medalists in both summer and winter. Besides, there is a positive correlation between the year and the number of participants. Based on these results, the ratio of medalists continued to decline until around 1940, and since then it has fluctuated between 20% and 30%.
- There is a positive correlation between the year of the Olympics and the number of events.

Proposal

- More attention should be paid to athletes between the age of 20 - 30 years old when selecting Olympic athletes. However, additional investigation is needed because the peak age is likely to vary by sport.
- It can be inferred that the acquisition of medals depends more on acquired factors than on innate factors such as motor skills. It would be beneficial to investigate the training methods of the U.S. or other developed countries to improve guidance, physical conditioning, and mindset for athletes to develop.
- Since the U.S. has by far the best results compared to other countries, a wide range of factors should be investigated, including training methods, training environment, culture, and so on.

Thank you!