SQL for Data Science Capstone Project

By

Cornelius Enetomhe

23/10/20245

The Project

The dataset

- The data used for this analysis is Olympics Dataset - 120 years of data. The dataset, originally, is made up of two files . athlete_events.csv and noc_regions.csv . This dataset was chosen because it gives the records of Olympics medal awards for different categories of sport, the countries that participated in the games, names, age, medals etc. The dataset is useful for news agencies reporting about the different feats in the 120 years of Olympics games existence. The dataset can also be a wealth of information for countries aspiring to improve on their performance in the subsequent Olympics events

The Client

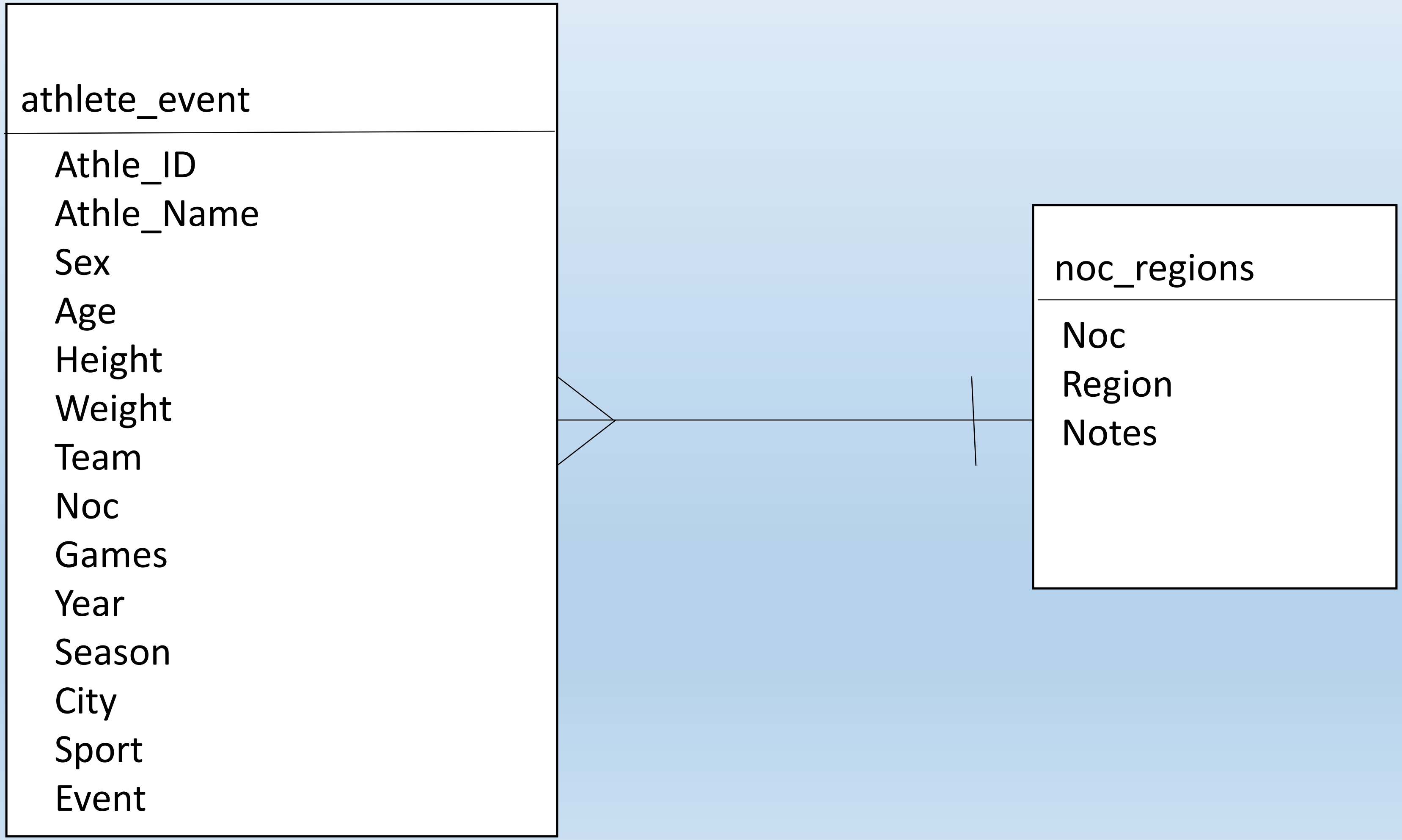
- SportStats, a sports analysis firm, looking to gain insights through trends ranging from the athletes to events from the Olympics dataset. The findings from the data will be shared to SportsStats partners, local news and elite personal trainers, for news stories and health insights. Not only will SportsStats and their partners will gain perspective but also those who are highly interested in sports or fitness.

PROJECT PROPOSAL

This project consists of showing data and statistics on the dominant countries in the Olympic Games, showing their development in the most popular sports, as well as showing which are the most popular events and competitions.

The result of this analysis seeks to help the company communicate key news to its partners by providing valuable information on sports trends.

PROPOSED ERD



athlete_event

- Athle_ID
- Athle_Name
- Sex
- Age
- Height
- Weight
- Team
- Noc
- Games
- Year
- Season
- City
- Sport
- Event

noc_regions

- Noc
- Region
- Notes

My major focus was on these questions:

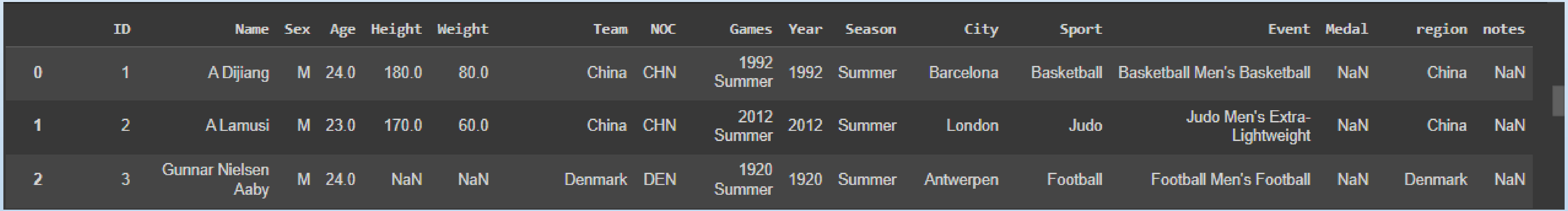
- Which countries have won the most gold medals in the Olympic Games?
- What is the most practiced sport?
- How many male and female events do we have?
- What is the relationship between age and sport?
- Does weight and height contribute to the award of a medal?
- Which country had the best ratio of participants to medals won?

Merge dataset

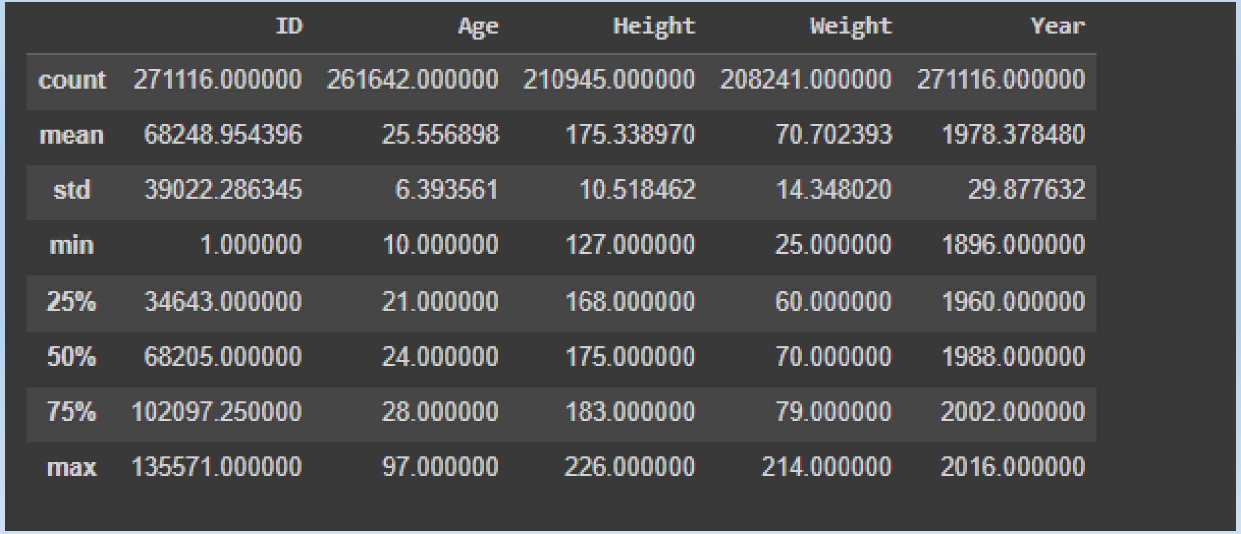
Importing the Datasets

To import data I made use of google.colab to gain access to the files on my google drive and mount it on my notebook and read it with pandas.

I combined together the athlete_events.csv and the noc_regions.csv files together. The merged table had 271116 rows and 17 columns

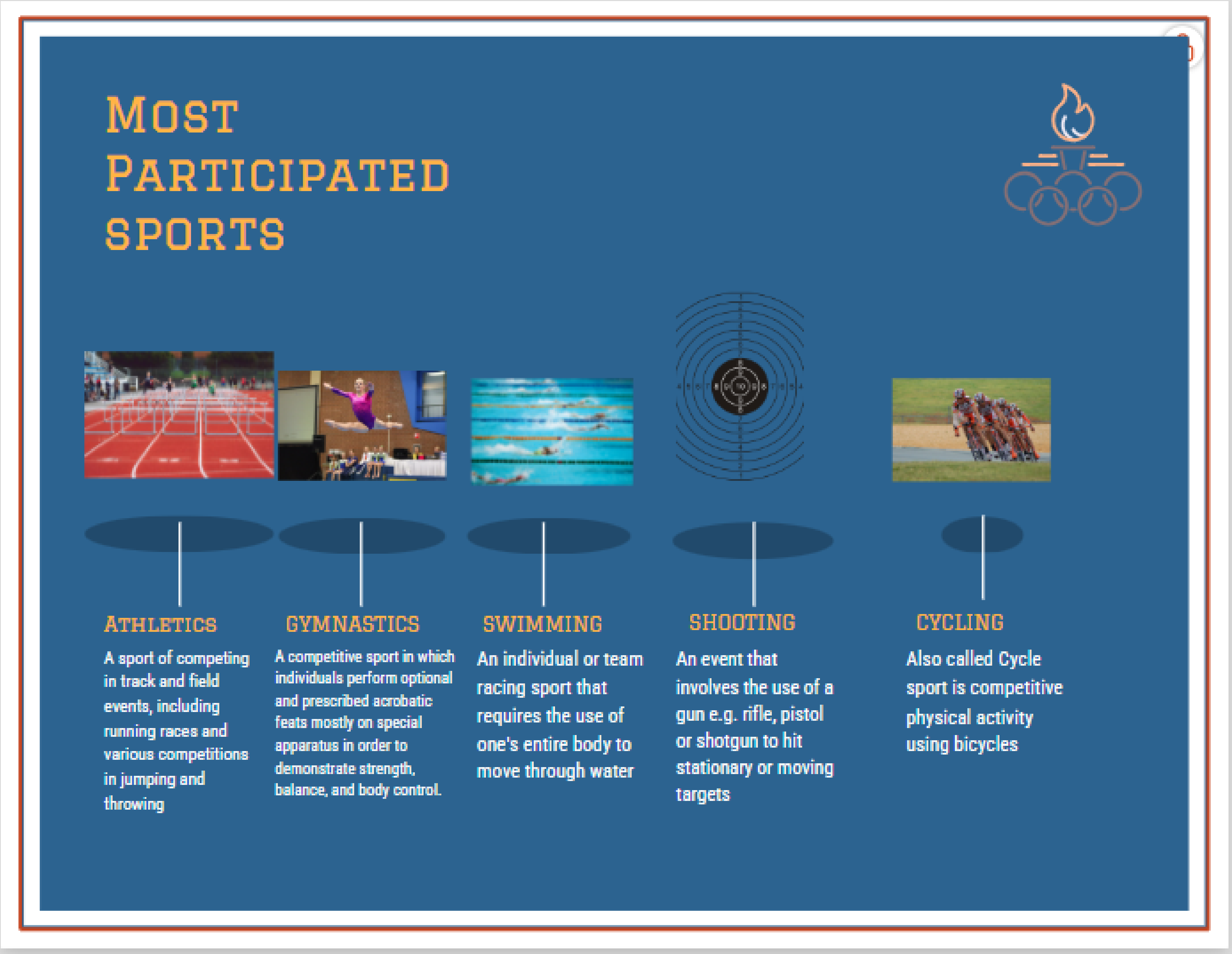


	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	region	notes
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN	China	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN	China	NaN
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN	Denmark	NaN



	ID	Age	Height	Weight	Year
count	271116.000000	261642.000000	210945.000000	208241.000000	271116.000000
mean	68248.954396	25.556898	175.338970	70.702393	1978.378480
std	39022.286345	6.393561	10.518462	14.348020	29.877632
min	1.000000	10.000000	127.000000	25.000000	1896.000000
25%	34643.000000	21.000000	168.000000	60.000000	1960.000000
50%	68205.000000	24.000000	175.000000	70.000000	1988.000000
75%	102097.250000	28.000000	183.000000	79.000000	2002.000000
max	135571.000000	97.000000	226.000000	214.000000	2016.000000

Basics summary of the dataset tells us the average ages, heights and weights of all participants in all events over the 120 years of the Olympic games.



Most Participated Sports

ATHLETICS

A sport of competing in track and field events, including running races and various competitions in jumping and throwing

GYMNASTICS

A competitive sport in which individuals perform optional and prescribed acrobatic feats mostly on special apparatus in order to demonstrate strength, balance, and body control.

SWIMMING

An individual or team racing sport that requires the use of one's entire body to move through water

SHOOTING

An event that involves the use of a gun e.g. rifle, pistol or shotgun to hit stationary or moving targets

CYCLING

Also called Cycle sport is competitive physical activity using bicycles

- Sporting events with the most number of participants over the Last 120 years of the Olympics.

Initial Hypothesis

- We would see that the countries that have more gold medals are those with that would have the most investment (USA, China).

- Factors such as height and weight are important but it also depends on the event they are participating in

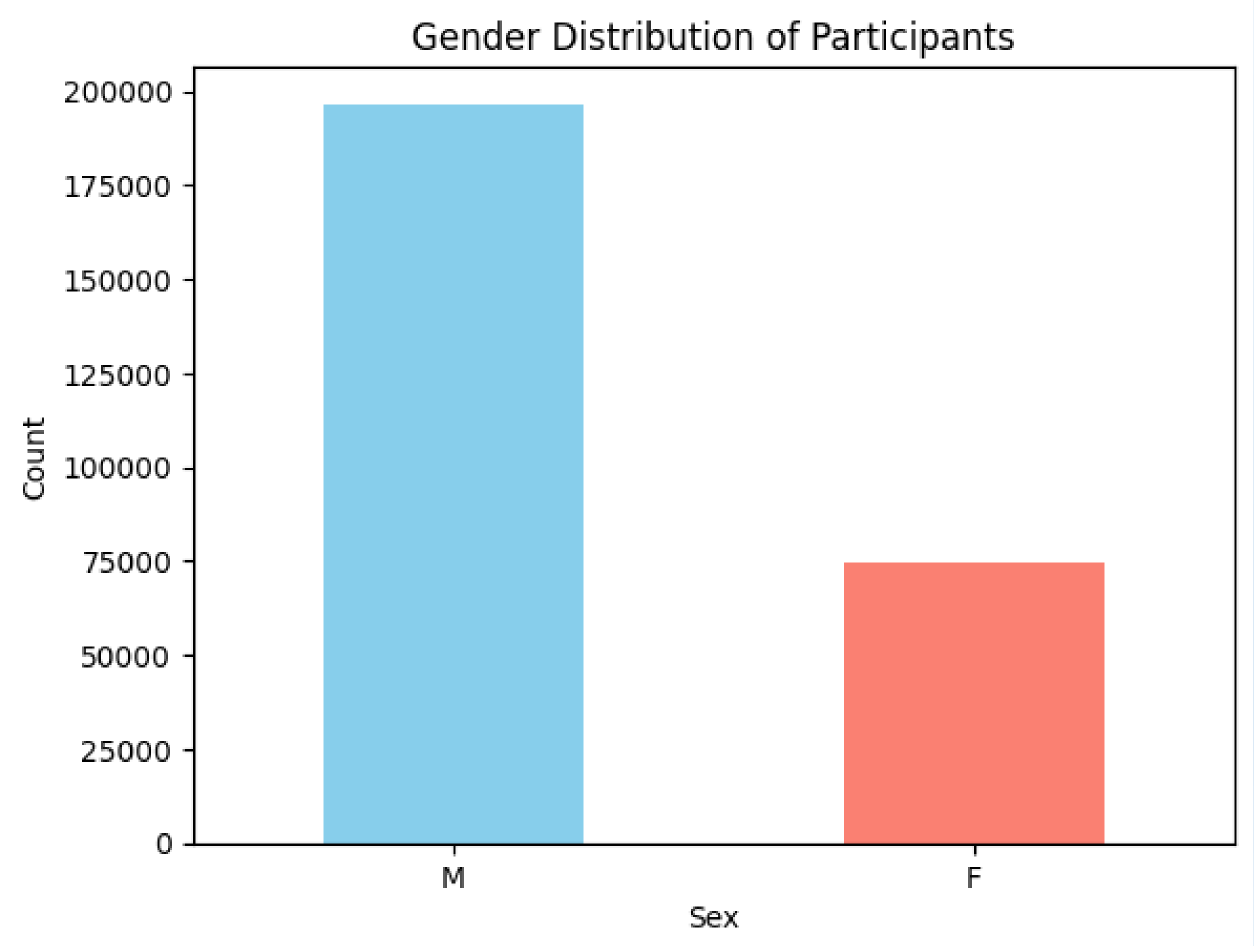
- The most participated sport would be track and field(athletics)

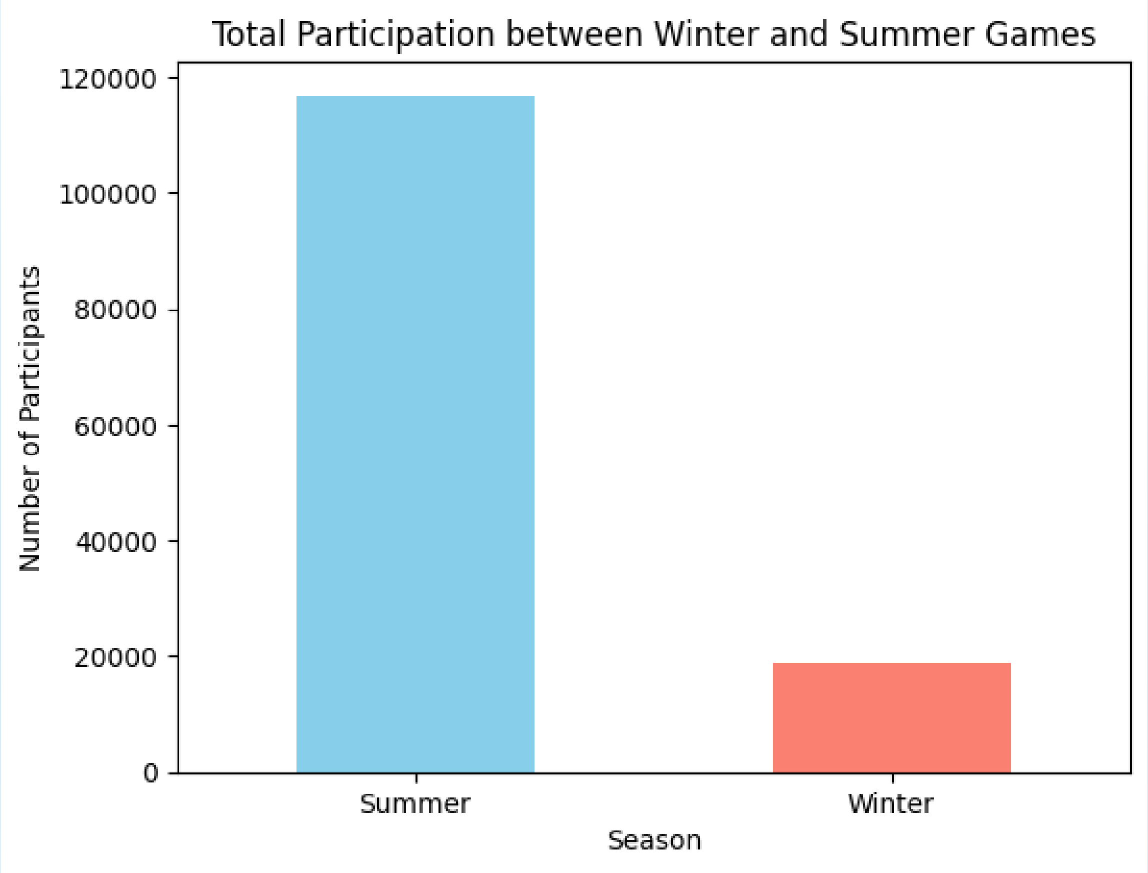
- These questions help guide my analysis and lead me to deeper analysis.

Statistics Summary

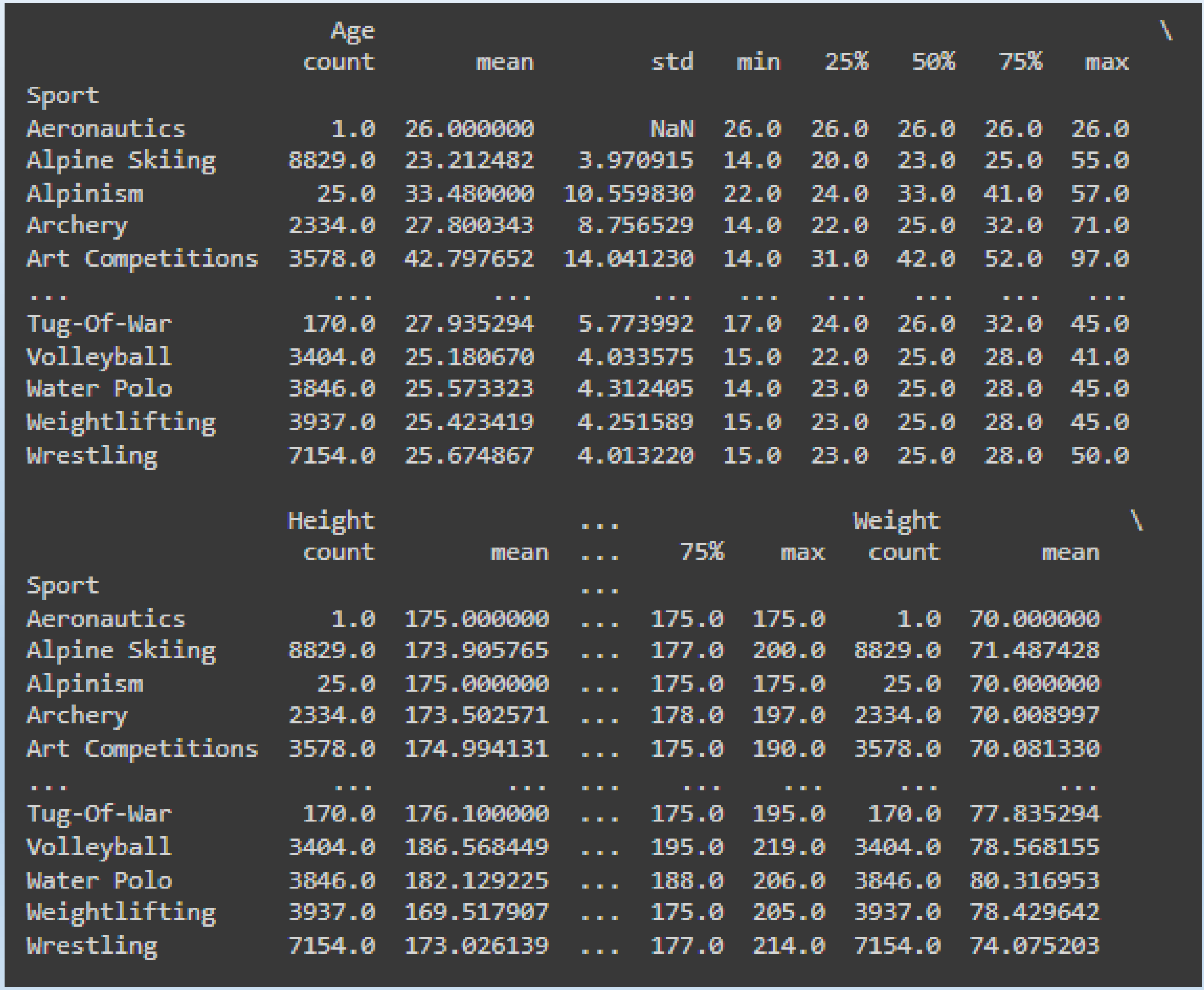
- We see a distinction between male and female participants over the years, mostly due to more male events.

- We also see the very small volume of winter events compared to that of the summer over the past few years



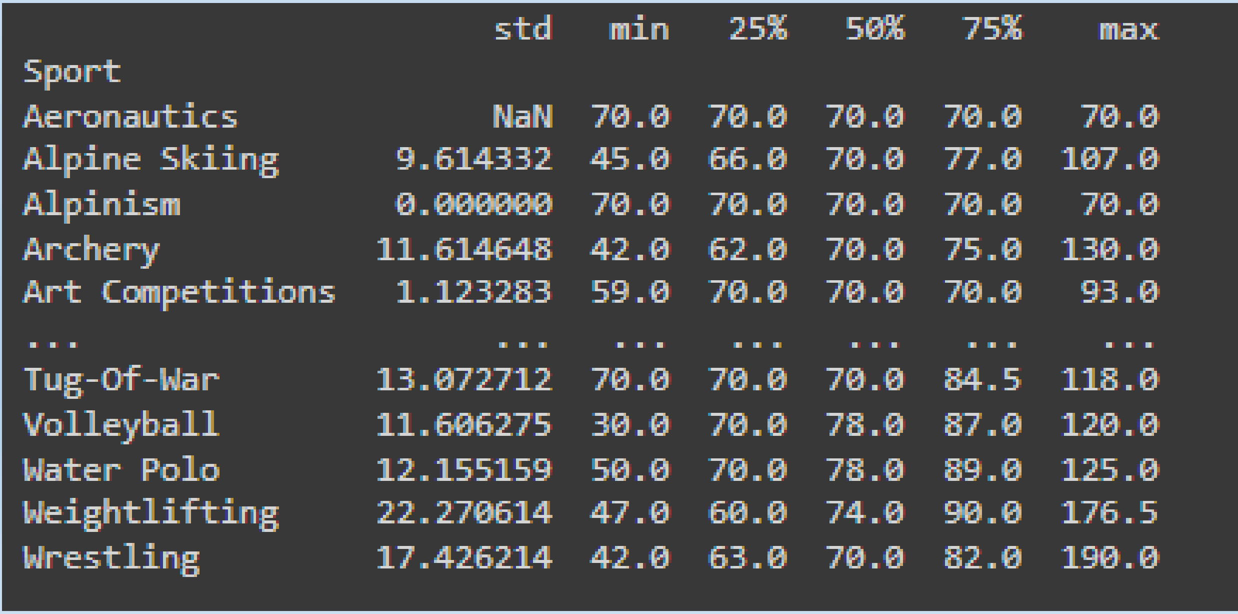


Age, Height and Weight distribution per sport



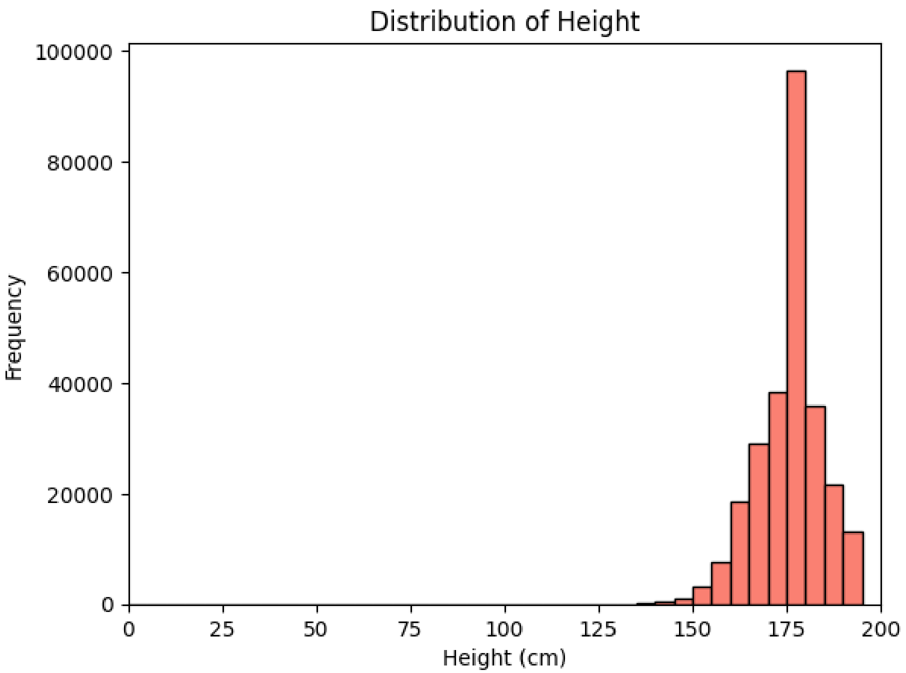
```
                    Age                                                     \
                  count       mean        std   min   25%   50%   75%   max
Sport
Aeronautics         1.0  26.000000        NaN  26.0  26.0  26.0  26.0  26.0
Alpine Skiing    8829.0  23.212482   3.970915  14.0  20.0  23.0  25.0  55.0
Alpinism           25.0  33.480000  10.559830  22.0  24.0  33.0  41.0  57.0
Archery          2334.0  27.800343   8.756529  14.0  22.0  25.0  32.0  71.0
Art Competitions 3578.0  42.797652  14.041230  14.0  31.0  42.0  52.0  97.0
...                 ...        ...        ...   ...   ...   ...   ...   ...
Tug-Of-War        170.0  27.935294   5.773992  17.0  24.0  26.0  32.0  45.0
Volleyball       3404.0  25.180670   4.033575  15.0  22.0  25.0  28.0  41.0
Water Polo       3846.0  25.573323   4.312405  14.0  23.0  25.0  28.0  45.0
Weightlifting    3937.0  25.423419   4.251589  15.0  23.0  25.0  28.0  45.0
Wrestling        7154.0  25.674867   4.013220  15.0  23.0  25.0  28.0  50.0

                 Height              ...                Weight             \
                  count        mean  ...    75%    max   count       mean
Sport                                ...
Aeronautics         1.0  175.000000  ...  175.0  175.0     1.0  70.000000
Alpine Skiing    8829.0  173.905765  ...  177.0  200.0  8829.0  71.487428
Alpinism           25.0  175.000000  ...  175.0  175.0    25.0  70.000000
Archery          2334.0  173.502571  ...  178.0  197.0  2334.0  70.008997
Art Competitions 3578.0  174.994131  ...  175.0  190.0  3578.0  70.081330
...                 ...         ...  ...    ...    ...     ...        ...
Tug-Of-War        170.0  176.100000  ...  175.0  195.0   170.0  77.835294
Volleyball       3404.0  186.568449  ...  195.0  219.0  3404.0  78.568155
Water Polo       3846.0  182.129225  ...  188.0  206.0  3846.0  80.316953
Weightlifting    3937.0  169.517907  ...  175.0  205.0  3937.0  78.429642
Wrestling        7154.0  173.026139  ...  177.0  214.0  7154.0  74.075203
```

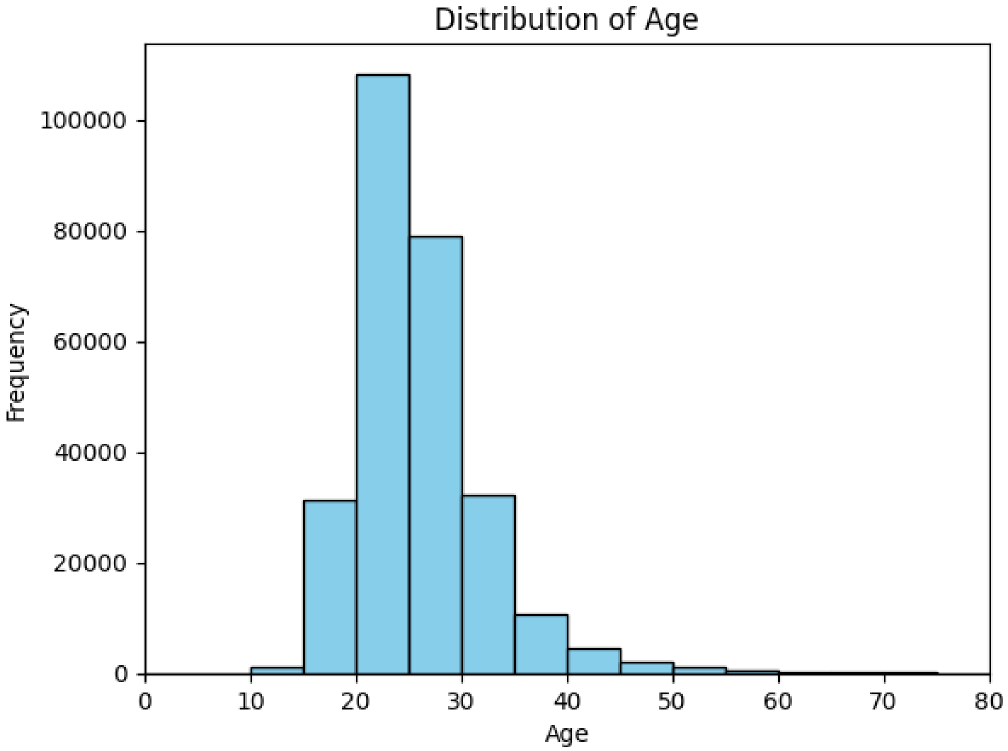
A view of the statistics summary of the distribution of Age, Height and Weight in relation to the various sporting events at the Olympics.

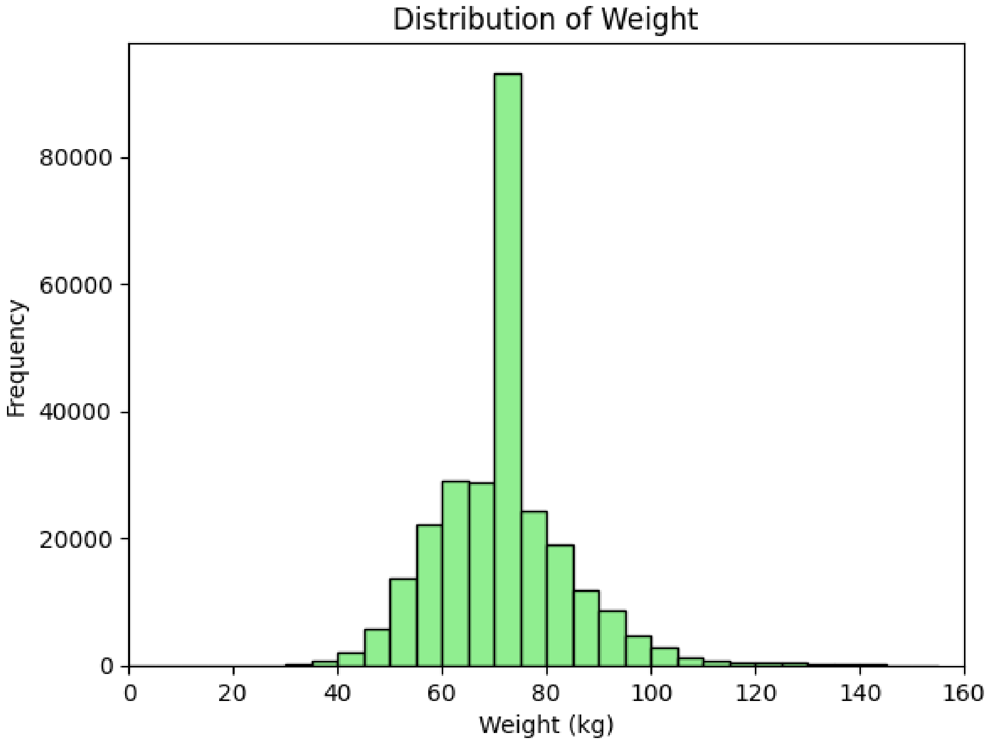


```
                        std   min   25%   50%   75%    max
Sport
Aeronautics             NaN  70.0  70.0  70.0  70.0   70.0
Alpine Skiing      9.614332  45.0  66.0  70.0  77.0  107.0
Alpinism           0.000000  70.0  70.0  70.0  70.0   70.0
Archery           11.614648  42.0  62.0  70.0  75.0  130.0
Art Competitions   1.123283  59.0  70.0  70.0  70.0   93.0
...                     ...   ...   ...   ...   ...    ...
Tug-Of-War        13.072712  70.0  70.0  70.0  84.5  118.0
Volleyball        11.606275  30.0  70.0  78.0  87.0  120.0
Water Polo        12.155159  50.0  70.0  78.0  89.0  125.0
Weightlifting     22.270614  47.0  60.0  74.0  90.0  176.5
Wrestling         17.426214  42.0  63.0  70.0  82.0  190.0
```

Deeper Analysis (Visualization of distributions)

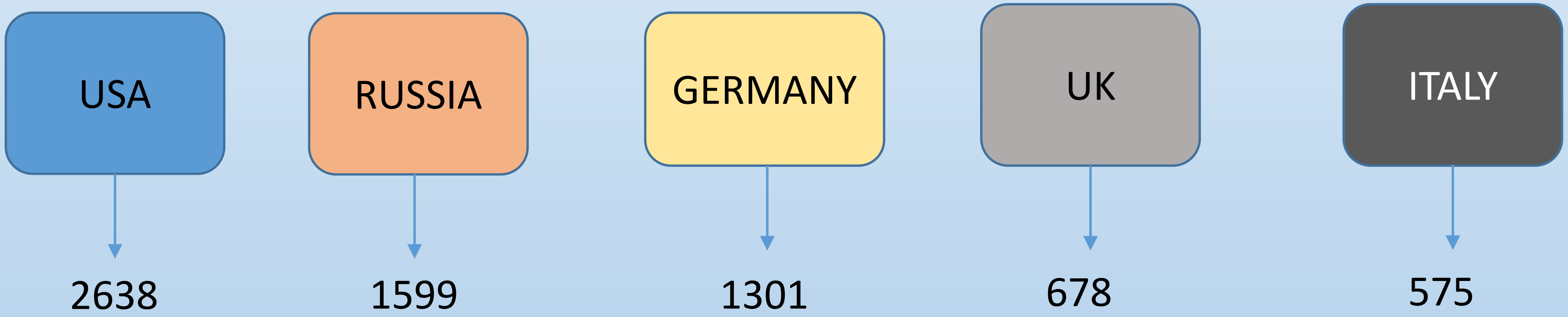






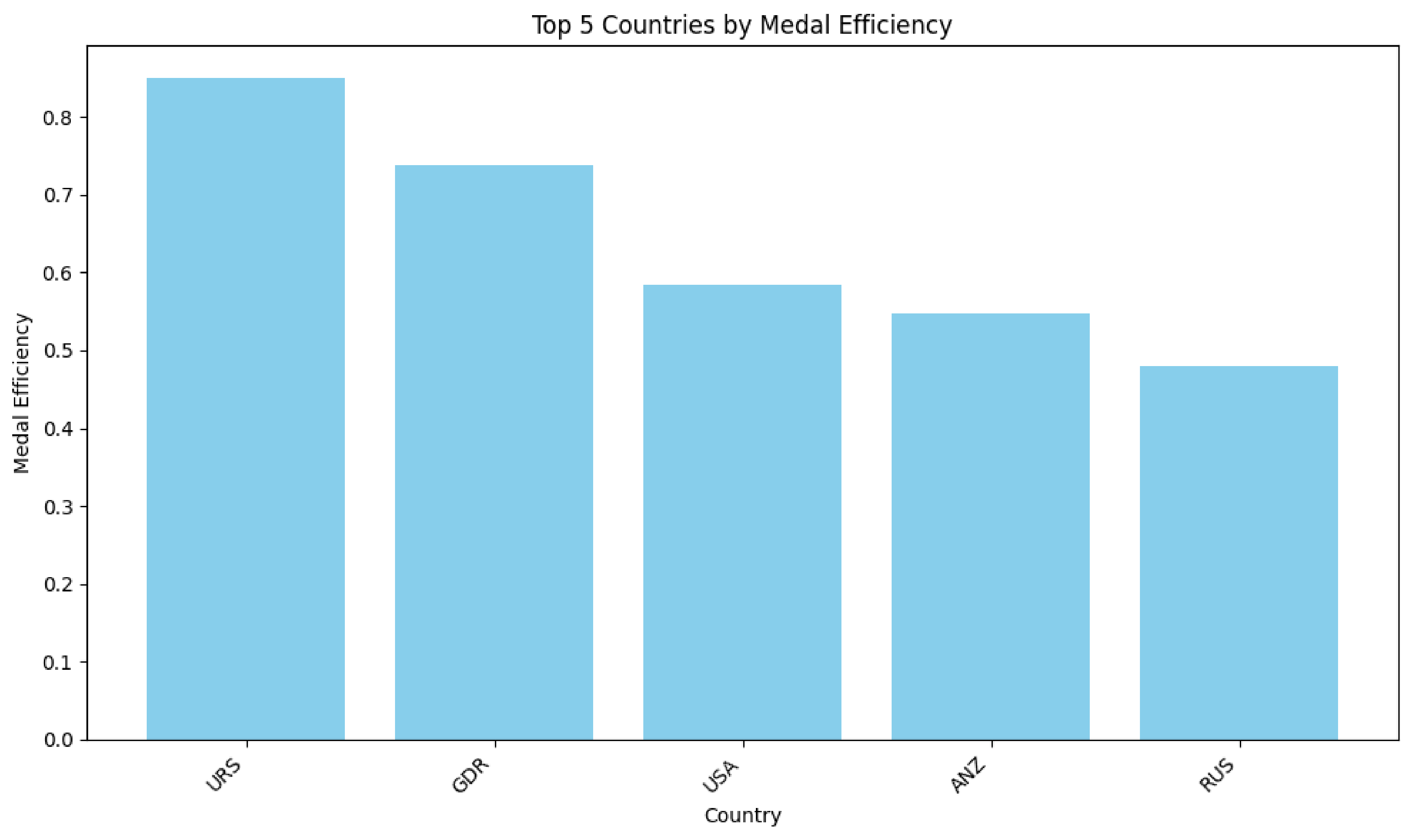
Deeper Analysis (Medals)

My initial hypotheses suggest USA and China as the top 2 based on investment. I was right about USA but China didn't make the top 5 for most gold but for total medals.



USA	RUSSIA	GERMANY	UK	ITALY
2638	1599	1301	678	575

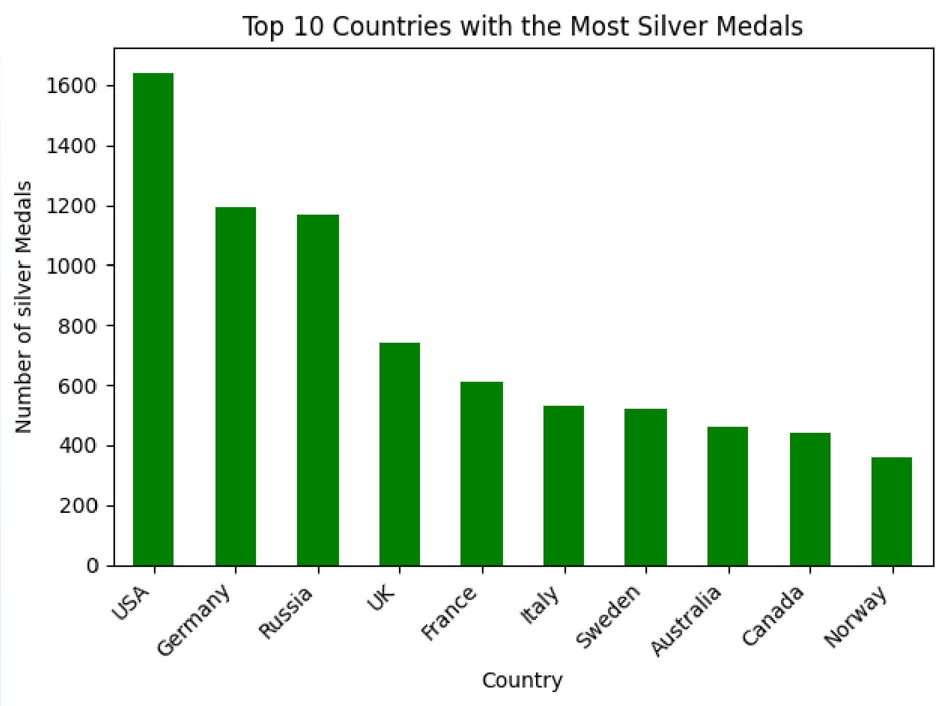
Deeper Analysis (Medals cont.)

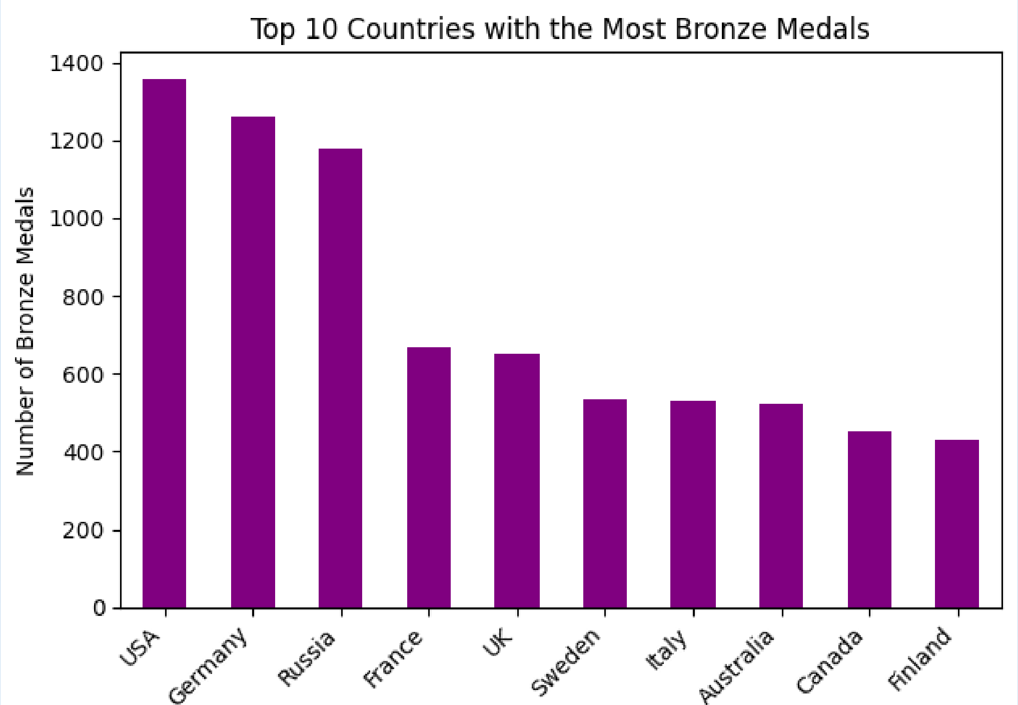


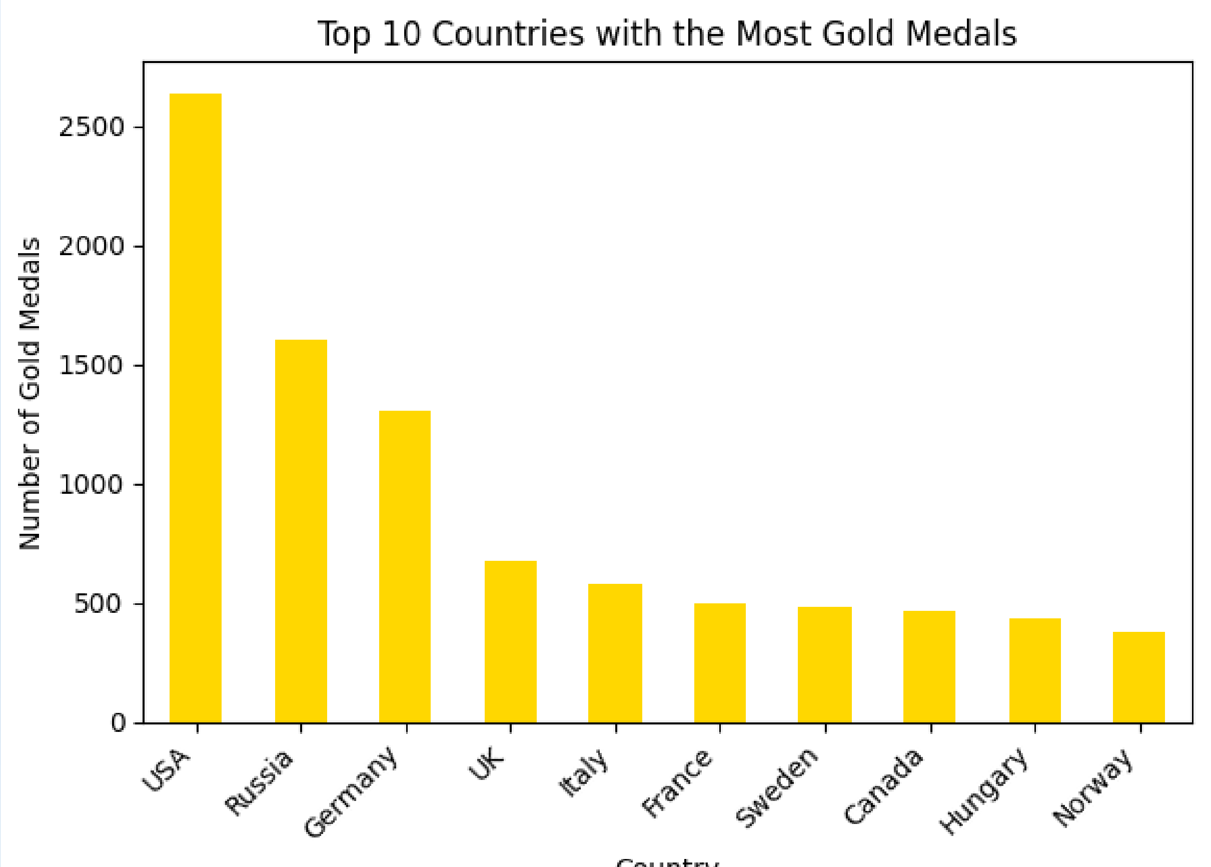
We see top 5 countries at the Olympics with the highest ratio of medals won to the number of participants representing them in all sporting events.

Deeper Analysis (TOP 10 Countries for Medals Categories)

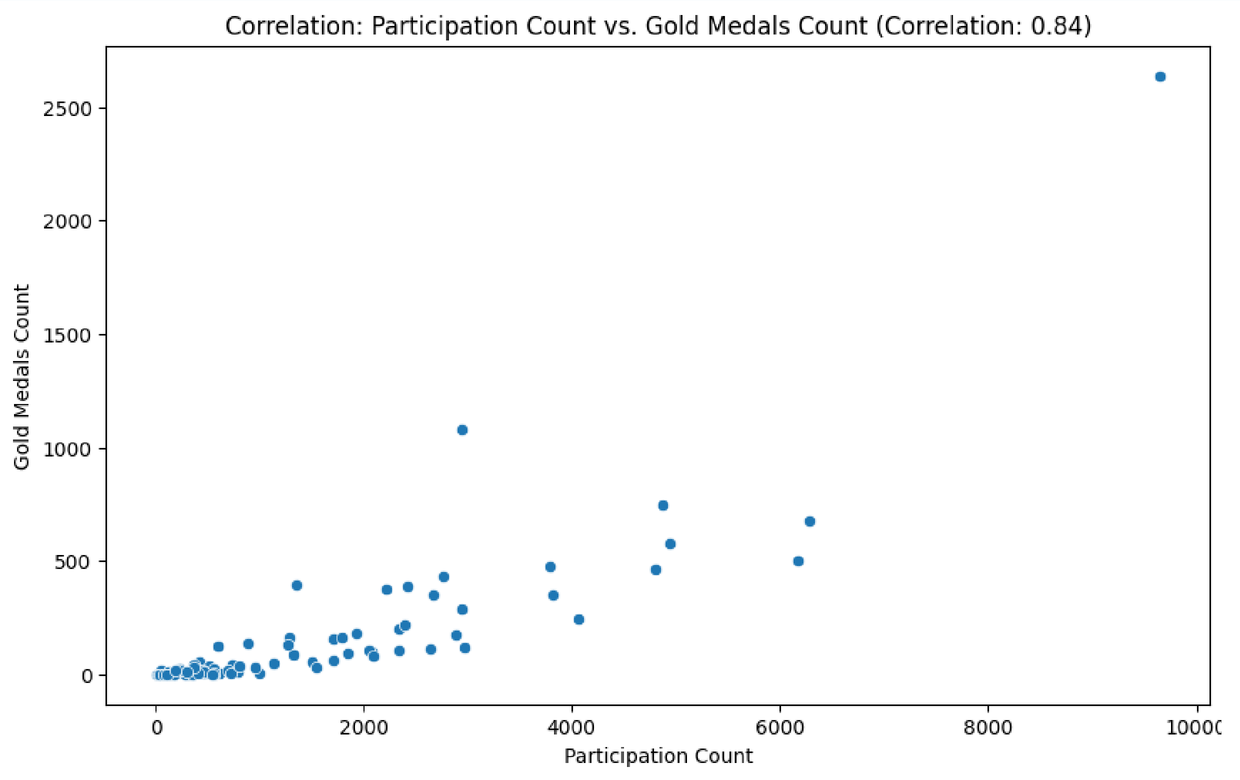
- We see countries like USA, Russia, Germany, UK , Sweden, Italy and France are in all categories for medals


Top 10 Countries with the Most Silver Medals

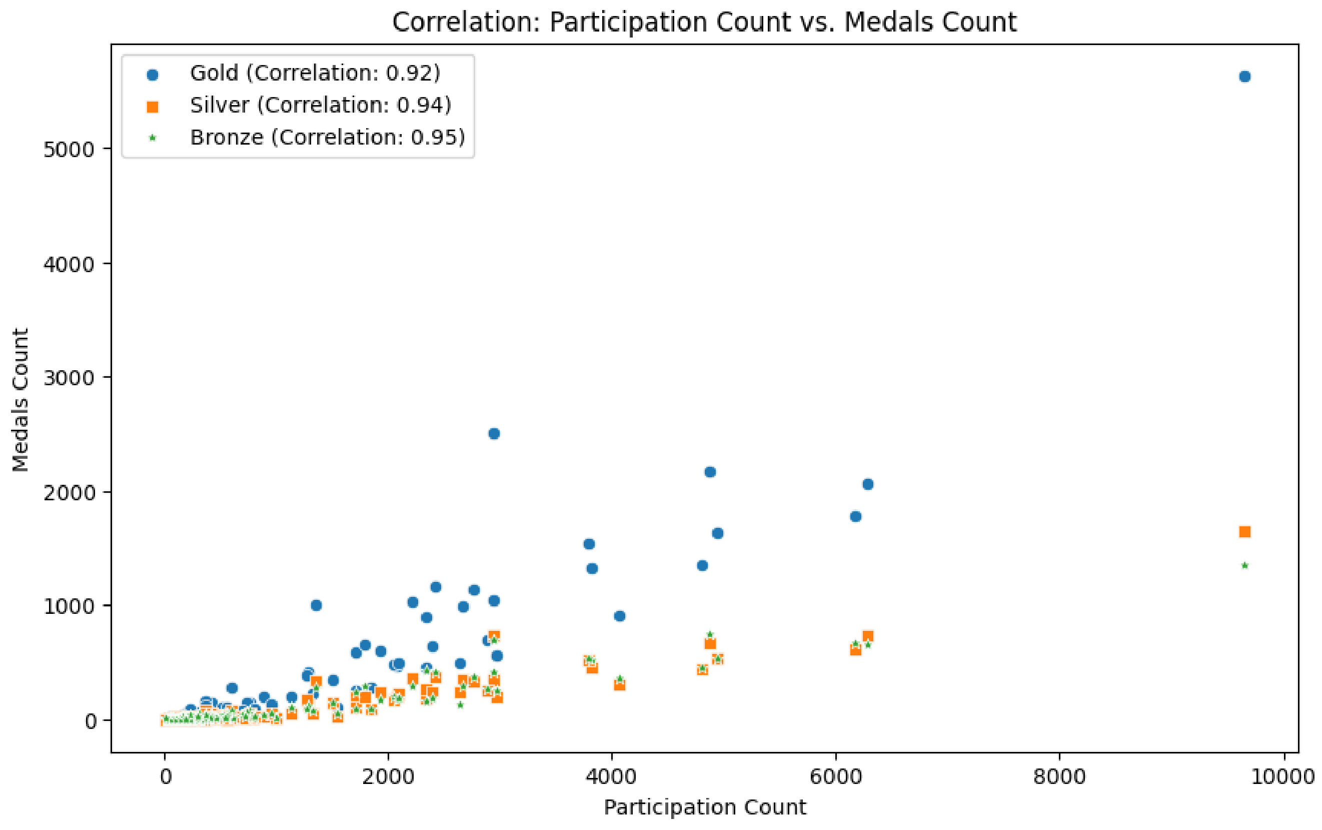

Top 10 Countries with the Most Bronze Medals


Top 10 Countries with the Most Gold Medals

Deeper Analysis(cont.)



In the participation by country and medals won Analysis, we can see that the correlation between Participation Count and Total Medals Won is 0.92.From this we can infer a positive linear relationship, because as participation increases due to more events or other factors so does the number of medals to be won.



Final Findings(Results of Hypotheses)

The following conclusions can be inferred from the Olympics data

1. We would see that the countries that have more gold medals are those with that would have the most investment (USA, China).
    - USA is the top performing country with both most medals and most gold medals and also ranks 3rd for medal efficiency
    - China didn’t feature amongst top 10 countries for most gold medals
    - Medals have linear relationships with participation and performance more than anything
    - I also wasn't able to prove or disprove if investment had an effect on medals won.
2. The most participated sport would be track and field(athletics)
    - We have seen that amongst the various events done in the olympics, athletics is the sport that has been done the most
3. Factors such as height and weight are important but it also depends on the event they are participating in
    - Each sport event did have different average ranges for age, height and weight.
    - Average height and weight also had difference by gender.

Recommendation

- Coaches can help guide prospective athletes by using average height and weight in various events to get the in best shape to win medals.
- Countries should do more to ensure their athletes pass the qualifying rounds to stand a better chance of winning
- Athletes should be encouraged to take part in qualifying rounds for the Olympics as early in their careers as possible
- Further analysis can be done on countries that have dominated some events and factors that contributed to that