“120 years of Olympic history: athletes and results” Dataset Analysis

The dataset used for this project can be found at the following link

Dataset Description

The data set contains information about the Olympic Games from 1896 to 2016. Each of the rows corresponds to the information of a specific athlete.

The **objectives** for this analysis are the following:

1. **Examine** and clean the data set
2. **Explore** numerical and categorical distributions utilizing statistics and graphs
3. **Explore** relationships of multiple characteristics using statistics and graphs

The data set has 271,116 rows and 14 columns. When performing a general analysis, it was discovered that several columns in the set had missing data, which could affect the final results.

The columns that presented missing data were:

- Age
- Height
- Weight
- Medal

Missing Values' Handling

As a first task, we will have to treat the null data in the set by filling those cells with information that does not affect the final result to be obtained.

To fill in the null data in the numerical columns, I made the decision to use the average of those columns. Here is my justification.

If we have the following Data Series:

```
[1, 2, NaN, 4, 5]
```

When calculating the average of that series with Python, we obtain the following result: 3.0. This is because, when calculating an average of a series with null data in it, Python does not take it into account and simply does the operation with the remaining data, in this case 1, 2, 4, 5. By adding them and dividing by the length of the series, we obtain the result 3.0.

It is important to process these missing data in the data collections that we are working on since, if we do not do so, at the moment of continuing with our analysis we could lose information on the rows where these null data is found.

For this reason, we are going to look for the best way to fill that missing data with usable values.

The first thing we should look for is to fill that missing data with a value that does not alter the results of the different operations that can be done on that collection of data. For example, what would happen if we changed the null values with zeros (0)?

```
[1, 2, 0, 4, 5]
```

If we do that and recalculate the average, we will see that we now have the following result: 2.4.

This result is wrong, since now, Python takes the entire collection of data, including zero, and divides it by the length of the collection, which went from 4 to 5. Adding the zero does not change the result of the sum, but it does change the length of the Series with which the division of the average will be done, so using zero as a change in the null data is not the best option.
Another way to change that null data to a value that can be used and that does not alter the collection like zero does, is to fill it with the average of the original collection. For example, we know that the average of the collection we have been using to do the explanation is 3.0.

```
[1, 2, 3, 4, 5]
```

If we use that average as a substitute of the null values on the series, the average of this new collection will continue to be 3.0, with this we do not alter the result that we can obtain from that collection.

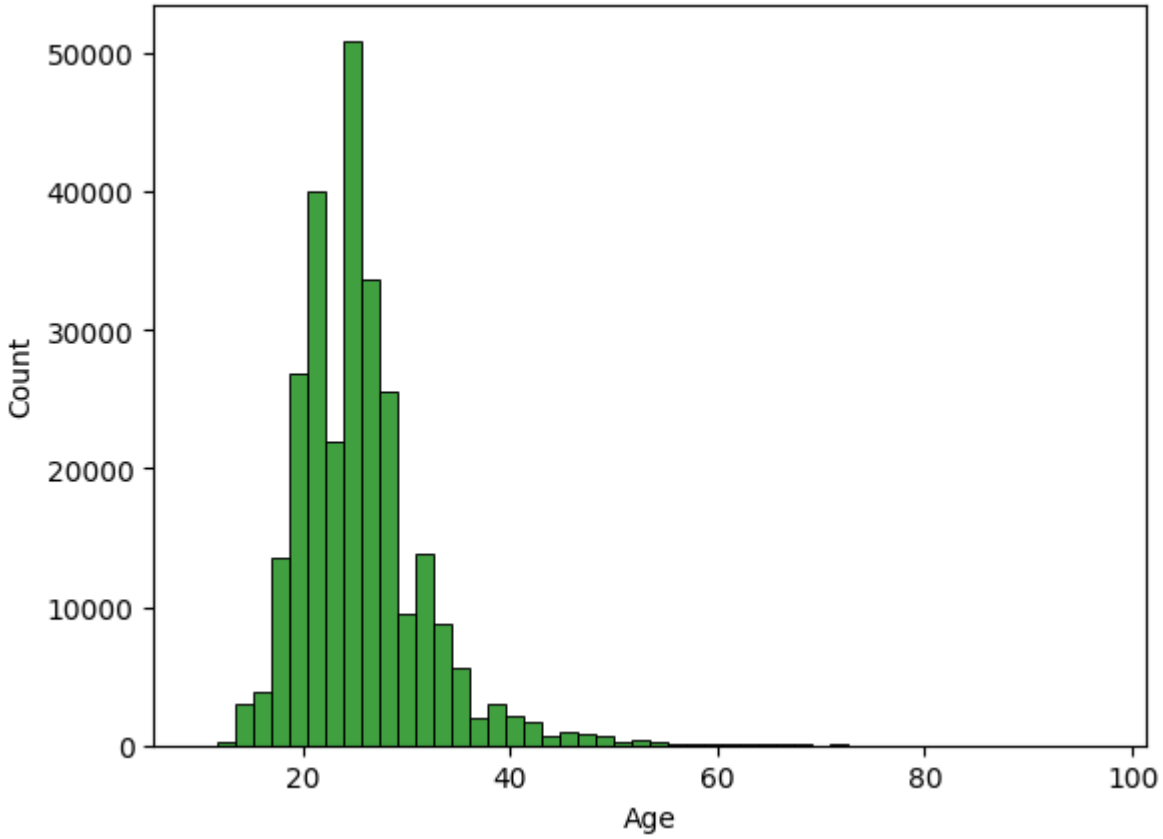
For the medal column that stores text, we will simply change the null values to the text "No medal info".

In this way we ensure that all columns have the same amount of data (271,116)

Exploratory Data Analysis (EDA)

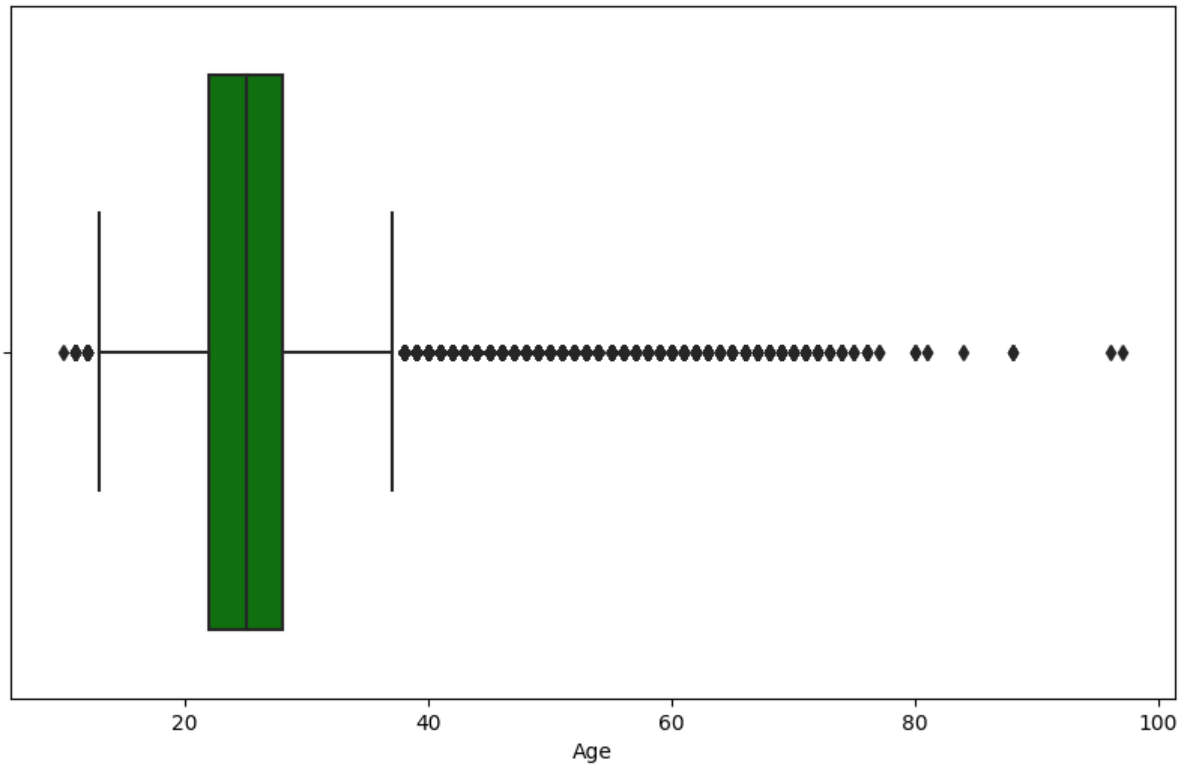
The next thing I will do on the dataset will be to obtain a series of graphs that will help us have an overview of it, to know which path to take when conducting our analysis.

Firstly, we will draw a histogram with which we will see the age distribution of the athletes.



We can see that the vast majority of the values are between 20 and 40 years old, although the histogram extends up to 100 on the x-axis, which means that in this dataset there are very old athletes who participated in the Olympic Games.

To better understand the data distribution, we will draw a Boxplot to see how these values are distributed.



The boxplot gives us different statistical summaries that help us understand the distribution of the data. Among them we can find the following:

- The box contains most of the data. In this box we can find 50% of the data in our set. We can check this by just looking at the histogram.
- The lines or "shovels" on both sides of the box represent the range the data has to be in order not to be considered atypical . In this case, they have a value of 1.5 * IQR (Interquartile Range). In this graph, the lines correspond to the following values:

    - 28 + (1.5 * 6) = **37**
    - 22 - (1.5 * 6) = **13**

    Any data that is above 37 or below 13 will be considered atypical.

Now that we know the values where ages are considered atypical, we will filter by those ages to see more specific information. With this filtering, we will find the following:

- The data entries of athletes older than 37 years old are **11,928**
- The data entries of athletes younger than 13 years old are **53**

We now know that there are very old and very young athletes who have participated in the Olympic Games, so now we will look for more information about those sports in which these athletes participated.

Young Athletes

Sport	***Number of Athletes***
Swimming	25
Figure Skating	15
Rowing	5
Gymnastics	5
Athletics	2
Diving	1

Old Athletes

(Here I decided to show the 5 most popular categories due to the fact that the 11,928 athletes were divided in +50 different disciplines, and showing each one of them is not useful for the conclusions I want to make from this data)

Sport	***Number of Athletes***
Shooting	3178
Art Competitions	2226
Equestrianism	1997
Sailing	1040
Fencing	1031

With this information we can notice a pattern that is repeated, and that is that younger athletes tend to practice sports in which the Physical Condition is a very important aspect (such as swimming or rowing) while older athletes lean towards disciplines where other factors such as Mental Acuity or Strategy are more relevant (such as art or fencing competitions)

Information in the different categories

Now I am going to look for more specific information within the dataset, and the first step to take is getting to know the distribution of the different variables.

Men and Women's Distribution

Sex	***Quantity*** ***(total)***
Men	196,594
Women	74,522

Countries with the most athletes

(Referring to all the athletes that have made it to their country's Olympic Team, not the total of athletes that a country took to a single edition of the OOGG)

Country	***Quantity***
United States	17,847
France	11,988
Great Britain	11,404
Italy	10,260
Germany	9,326

Year and Season that had the most athletes

Season	***Quantity***
2000 Summer	13,821
1996 Summer	13,780
2016 Summer	13,688
2008 Summer	13,602
2004 Summer	13,443

Number of athletes in different seasons

Season	***Quantity*** *(total)*
Summer	222,552
Winter	48,564

City that has received the most athletes

(In total, not in a single edition of the OOGG)

City	***Quantity***
London	22,426
Athenas	15,556
Sidney	13,821
Atlanta	13,780
Rio de Janeiro	13,688

Sports with more athletes

(In total, not in a single edition of the OOGG)

Sport	***Quantity***
Athletics	38,624
Gymnastics	26,707
Swimming	23,195
Shooting	11,448
Cycling	10,859

Events that most athletes have had

(In total, not in a single edition of the OOGG)

Event	***Quantity***
Men’s Football	5,733
Men’s Ice Hockey	4,762
Men’s Hockey	3,958
Men’s Water Polo	3,358
Men’s Basketball	3,280

Medal Info

Medals	***Quantity*** *(total)*
No medal Info	231,333
Gold	13,372
Bronze	13,295
Silver	13,116

Athletes’ Information

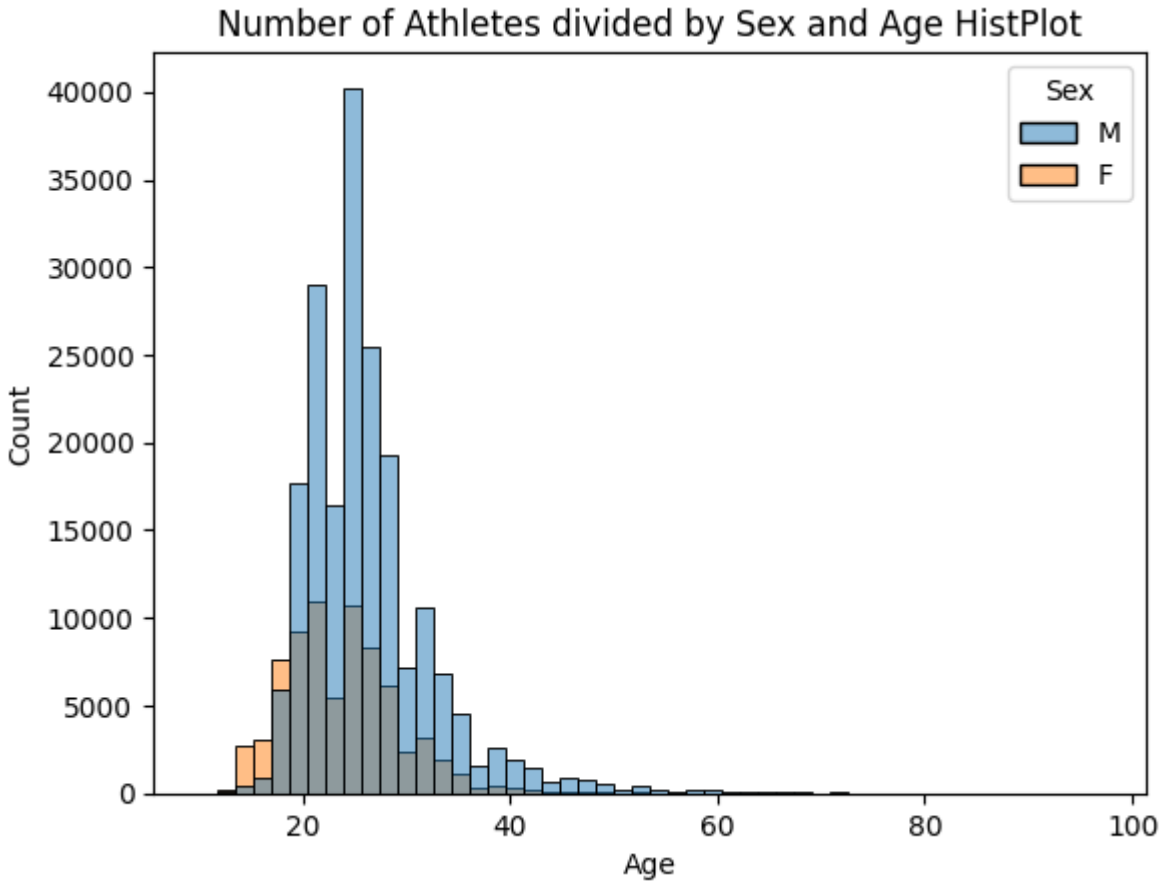
Now that we know more about the dataset and the distribution of the information within it, we can get back to continue analyzing statistical variables related to the athletes and conclude the analysis.

The next table shows us the average Age, Height and Weight of the athletes, divided by gender, throughout all the history of the Olympic Games:

Sex	***Age***	***Height*** *(CM)*	***Weight*** *(KG)*
F	24	168	61
M	26	178	74

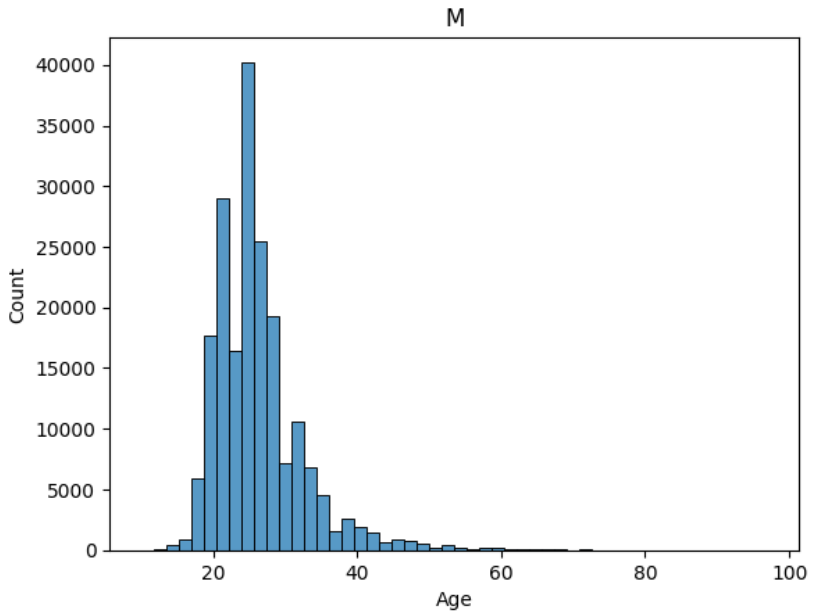
We can see that, at least in general, the men present at the Olympic games are taller and weigh more than the female athletes.

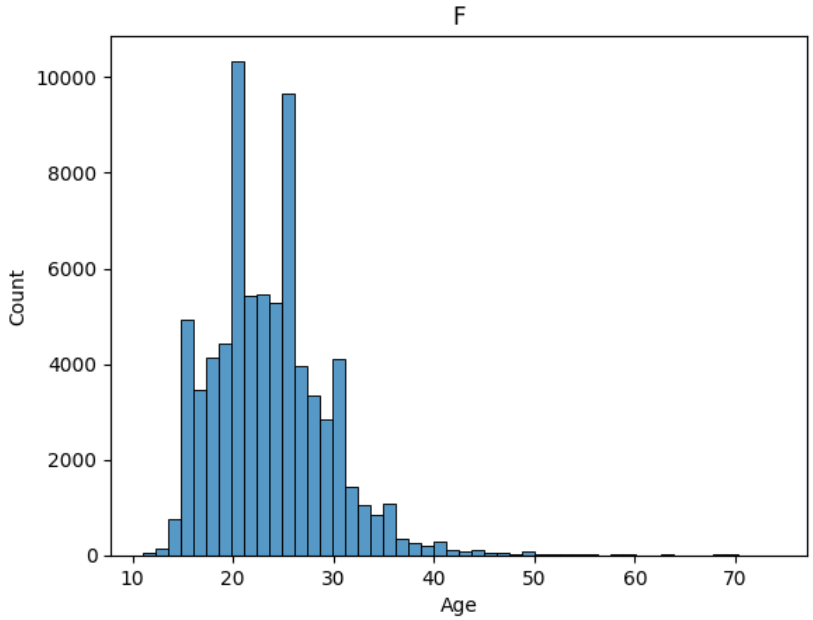
In the following histogram we can see the distribution of the number of female and male athletes, divided by gender and age, to show the disparity between men's and women's participation.



With this we can verify that historically there are no equal conditions, since men have had many more opportunities to compete than women in the Olympic Games.

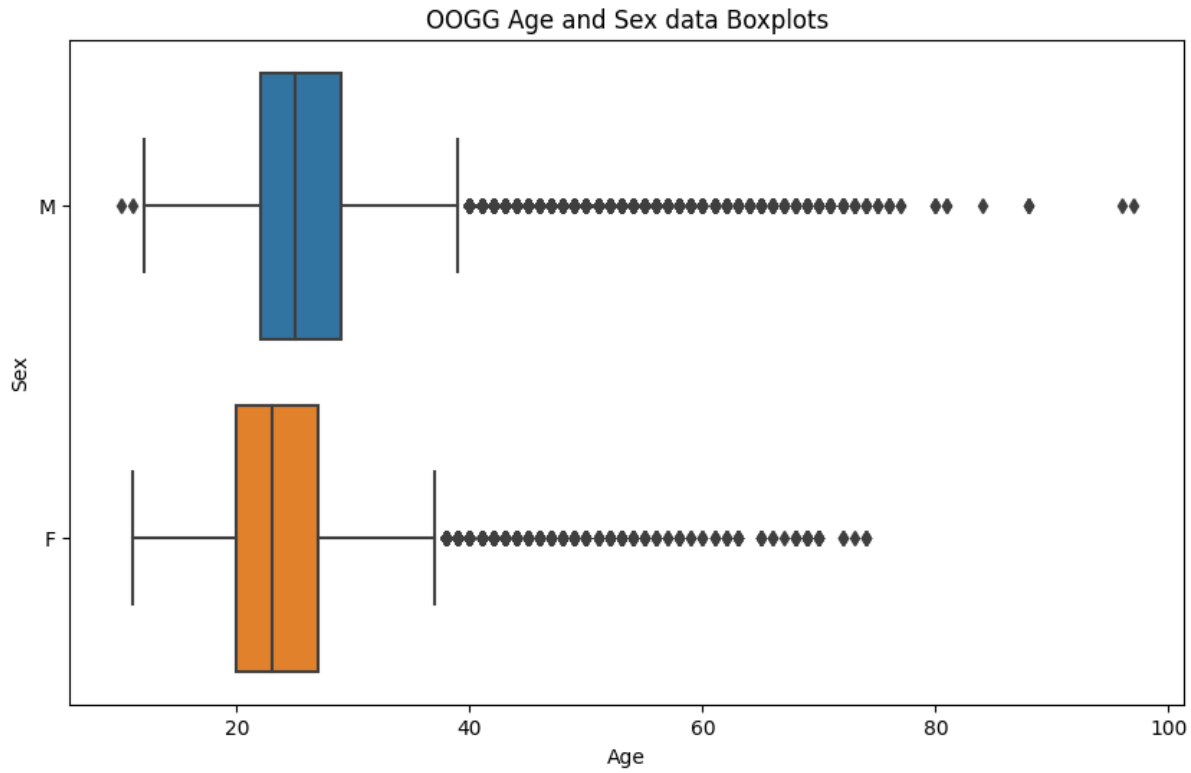
To better understand the distribution of men and women, we are going to divide the graph into two different ones.





With these two graphs we can see perfectly that much less women have participated in the OOGG.

And interestingly, we can see that while the women tend to start at an earlier age, the oldest athletes are usually men. Let’s confirm this with a boxplot.



Here we can observe two important points:

- The Women boxplot starts to draw earlier, and lower ages are considered typical comparing it with the Men one.
- Men tend to live longer and that is why their atypical values are greater.

Additionally, the following table shows the number of teams, sports and total events that men and women have had, divided by season.

Season	***Sex***	***Number of Teams***	***Number of Sports***	***Number of Events***
Summer	F	352	40	214
	M	1118	49	491
Winter	F	144	14	57
	M	214	17	67

We can see that the Summer Olympics are where there is the biggest difference between the number of men's and women's teams competing as well as in the events. And although the Winter Olympic Games have been around for less time, you can still notice this disparity in them.

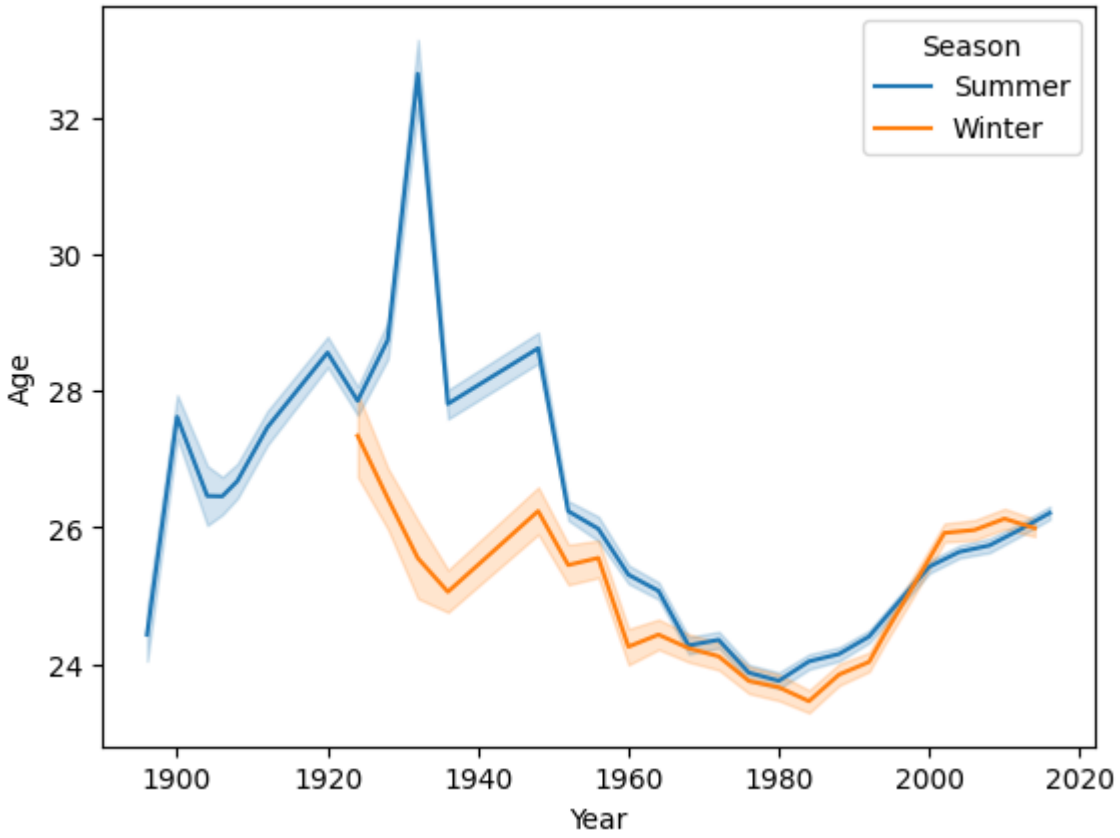
Before we finish, in the information on the athletes who win medals we can see that the average age, height and weight of all of them is very similar. Here is that information broken down:

Medal	***Season***	***Sex***	***Age***	***Height*** *(CM)*	***Weight*** *(KG)*
Gold	Summer	F	24.21	171.67	64.38
		M	26.47	179.87	76.87
	Winter	F	25.20	167.62	62.43
		M	26.60	179.54	77.78
Silver	Summer	F	24.29	171.39	64.06
		M	26.63	179.48	76.49
	Winter	F	25.24	167.97	62.26
		M	26.43	179.09	77.25
Bronze	Summer	F	24.63	171.18	64.05
		M	26.32	179.43	76.39
	Winter	F	25.12	167.40	61.38
		M	26.38	178.89	77.10

This table tells us the following:

- The age range of the athletes who win Medals, therefore their “Physical Prime” is between the 24 and 26 years of age generally.
- The height of those athletes ranges between 167 and 179 cm.
- The weight of medal-winning athletes does not exceed 78 kg.

Lastly, knowing that the winning athletes are very young, we will make a graph to see how the average age of the athletes fluctuated over the years, and in the different seasons of the OOGG.



We see that over the years the average age of the athletes has been very stable and almost always remained well below 30 years, except in the period between the two World Wars (this could be due to the unfortunate death of many young people that were sent to fight for their country, this resulting in the recruitment of older athletes in order to complete the Olympic Teams)

Conclusions

In the Olympic Games, men have been the main figures in the different competitions. And this has not changed much over the years, since if we compare the number of male athletes against female athletes, we see a difference of 122,075 athletes in favor of men. In other words, male participation is 62.09% higher than female's, historically. Something important to keep in mind is that the first female athlete made her appearance at the Olympic Games in Paris in 1900, but almost 100 years later, in Barcelona in 1992, there were still 35 countries that only had men's teams.

Now, moving on to the age-related information, but not leaving aside the gender distinction yet, by analyzing atypical values, we can see that male athletes keep on practicing the sport much longer than their female counterparts (Men being the oldest athletes ever recorded who participated in the Olympic Games: 97 years for the summer season and 58 for the winter season), and that contributes to the inequality between the two.

Talking about outliers or atypical values, quite a lot of athletes are outside the "typical" age range of 13 to 37 years old. And what is interesting is that by analyzing this information and seeing what disciplines these atypical athletes choose, we can see how the physical demand of the chosen discipline is becoming less and less as the age of the competitor increases (Athletes under 13 choose very physically demanding sports, such as swimming or rowing, while athletes over 37 lean towards disciplines that do not require as much physical effort, such as shooting or art competitions.).

By continuing with this age analysis, we notice that the majority of athletes who participate in the Olympic Games are not older than 28 years old, and we can see a relationship of age with different data from the dataset, as the following below:

- The sports that most athletes participate in are disciplines where the physical effort required is very high, being the most popular Athletics, Gymnastics, Swimming, Shooting and Cycling.

- The average age of medal-winning athletes does not exceed 26 years, which indicates that in general, athletes are in their best physical condition to compete in the games during that age range.

Also, the average age has fluctuated along the years, but not for “athletical” reasons.

We saw that the average age of athletes since the beginning of the Olympic Games was always between 24 and 26 years old, but a historical event (or rather 2) altered these numbers for a good part of the 20th century.

First, the average age of athletes suddenly rose, reaching a peak of 33 years around 1930 (this coincided with the world recovering from the First World War, in which many young people died fighting for their country, so older athletes had to participate in the games). And second, During the post-World War II period, the average age fell sharply to reach its historical minimum of 23 years (This could be due to the fact that, contrary to what happened in the first war, now the athletes were people born after this tragic event ended).