An Analytical Study on Olympics

Foundations of Data Analytics
Project Report

Done By:-
Vinu Kevin Diesel .S.P(19BCE1194)
Revanth Bhargava Boyidi (19BCE1177)

Under the Guidance of:-
Dr. Tulasi Prasad Sariki

TABLE OF CONTENTS

Introduction

The modern Olympic Games, sometimes known as the Olympics, are the most important international athletic event, including contests in both summer and winter sports. Thousands of participants or players come from around the world to participate in a variety of competitions. Around 200 nations participate in the Olympics every time, making it the most famous sports competition. The Olympic Games are held every four years, with the summer and winter Olympics switching places every two years.

From 1896 to 2018, the Olympic Games were held. It would be an interesting topic to analyze the trends and patterns which it follows throughout history. We will bring forth the factors affecting the number of medals won, the trends in BMI of athletes on each game, the difference between winter and summer games and much more.

The data collected for this paper is from kaggle.com. It is about the Olympics that has taken place for 120 years, ranging from 1896 to 2016. The dataset consists of different characteristics of a participant like a name, gender, age, height, weight, team, medals won, sport, year of participation, from 1896 to 2016.

1. Overview

The Summer Olympic Games are an international sports competition hosted once in four years. Greece was the first country to host the Olympics in 1896 in Athens, and most recently, Japan is hosting it in Tokyo. The International Olympics Committee(IOC) organizes the Olympic Games in the selected host city. Different cities host the Olympics every four years.

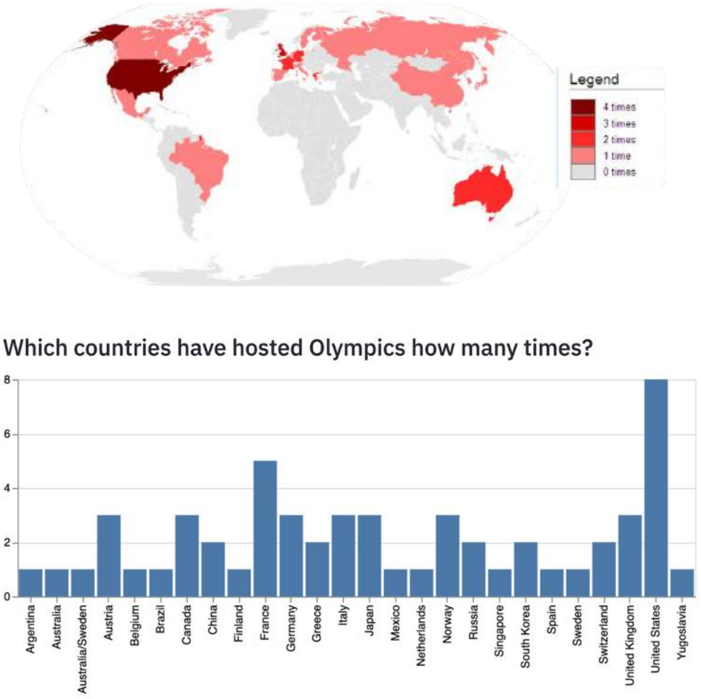


Figure - 1: World Map showing the countries and the number of times they have hoisted the Olympics

The success of the Summer Olympics gave birth to Winter Olympics in 1924. Summer and winter Olympics are held alternatively every two years from 1994. While around 204 countries participate in Summer Olympics, only 88 countries participate in Winter Olympics. The Summer Olympics have 300 athletic events across 28 sports, while the Winter Olympics feature 102 sporting events across 15 sports. And, there is a noticeable difference between the total medal boards of the Summer and the Winter Olympics.

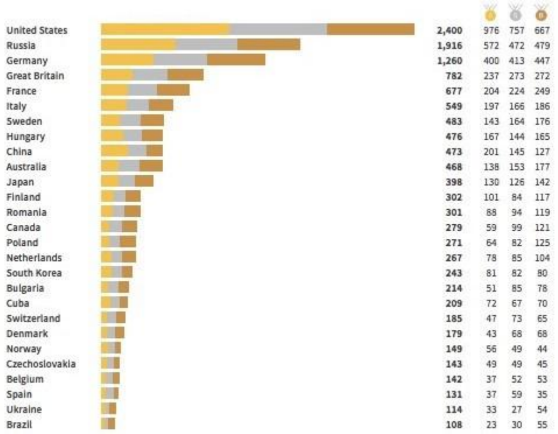


Figure - 2: Total Medal count of Summer Olympics

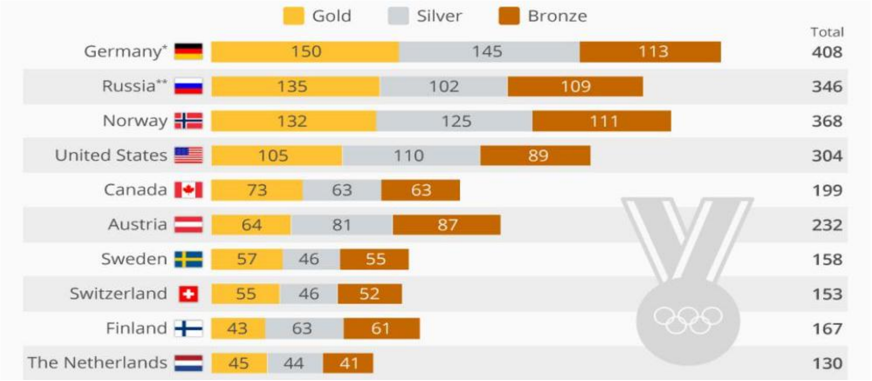


Figure - 3: Total Medal count of Winter Olympics

End of the day, the effort put in by the sportsperson results in a medal for the country. A sportsperson has to go through a lot of exercise, fitness, practice and others. These will have a high impact on their victory. But, to what extent does a sports person's bodily characteristics affect a country's victory?

The hosting country has lots of benefits like an increase in GDP, total revenue, so on. But, does the hosting country has an impact on its victory that year. The United States has hosted the Olympics eight times and is present at the top of the leader board. Is this a coincidence or not?

Problem statement

The Olympic Games are a well-known sports event that is well-known around the world. It has a long history that dates back to 1896 and has been a part of history. It's fascinating to observe how historical events have influenced the details of the Olympics and how they've changed through time.

Hence our project tries to answer these three main questions:-
1. What impact does the host country have on the number of Olympic medals won?
2. What are all the differences between summer and winter games?
3. Is there any patterns in the BMI of athletes who won certain games in the Olympics?

Objective and Scope of the Project

1. Objective

- Study the Olympics games data for finding dependency of the host country on the medals won using R studio

- Identifying various factors which differentiates summer and winter Olympics games using R studio

- To diagnosticate the relations between BMI of the athletes who won and specific game using graphs, done using Tableau

2. Scope

- The scope of the study covers all the countries which have been participating in the Olympics games

- The study covers 120 years of Olympics games data starting from 1896 to 2016

3. Out of Scope

- The dataset does not include details or participants of 2020 Olympic Games that are held in Tokyo.
- Most important factors for a player's victory, talent and practice are not considered for the analysis, as these cannot to collected or tabulated.

Data Source

This data was taken from the Kaggle repository. Originally this data was scraped from "www.sports-reference.com". This dataset has the details of all the participants and the events & the season they participated. This dataset also has the winner of each game with the medal won.

The columns of the dataset are:-

- **ID** - Unique number for each athlete
- **Name** - Name of participant
- **Sex** - M or F
- **Age** - Integer
- **Height** - In centimeters
- **Weight** - In kilograms
- **Team** - Team name

- **NOC** - National Olympic Committee 3-letter code
- **Games** - Year and season
- **Year** - Integer
- **Season** - Summer or Winter
- **City** - Host city
- **Sport** - Sport
- **Event** - Event
- **Medal** - Gold, Silver, Bronze, or NA

This data was collected and tabulated during May 2018 from the **Sports Reference** website. This website operates a group of sites providing basic statistics and resources for sports analysts, fans and others.

Tools & Techniques

We analyzed the data using the following analytical techniques and methodology:-

- Identify the central tendency for each characteristic.
- Univariate Analysis: Visualize each variable using plots and graphs.
- Bivariate Analysis: Identify the relationship between two variables.
- Identify factors affecting a sportsperson victory using correlation and regression methodology.
- R - Used for analysis and visualization.
- Techniques: Bar graph, Line graph, Pie Chart, Histogram, Correlation Heatmap

Analytical Methodology:-

- Data extraction from the data source

- Data wrangling

- Study each variable and its significance to the study.

- Exploratory Data Analysis

- Each variable is subjected to a univariate analysis.

- Performing bivariate analysis between two variables

- Observations should be documented with sufficient evidence.

Limitations

- Medals column has many multiple values. Those rows can’t be dropped or replaced with mean, median, or mode.

2. Data Description and Preparation

Data Management

The dataset which was collected from the Kaggle repository contains almost all players who participated in the Olympic games from 1896 to 2016. Along with the name of a player, their gender, age, height, weight, country, sport played are given.

List of Variables And their Types		
Variable	**Description**	**Datatype**
ID	Unique number for each athlete	Continuous
Name	Name of participant	Discrete
Sex	M or F	Discrete
Age	in years	Discrete
Height	In centimetres	Continuous
Weight	In kilograms	Continuous
Team	Team name	Discrete
NOC	National Olympic Committee code	Discrete
Games	Year and season	Continuous
Year	Integer	Continuous
Season	Summer or Winter	Discrete
City	Host City	Discrete
Sport	Sport	Discrete
Event	Event	Discrete
Medal	Gold, Silver, Bronze, or NA	Discrete

Data Quality

- Height and weight for some of the players are missing. These records can't be removed as they will have a huge impact on the medal count of a country.

- Height and weight were collected from online sources, so their validity cannot be known. And these factors could vary from time to time.

Data Preparation

- Missing values in height and weight was filled with the average height and weight of players playing the same sport.

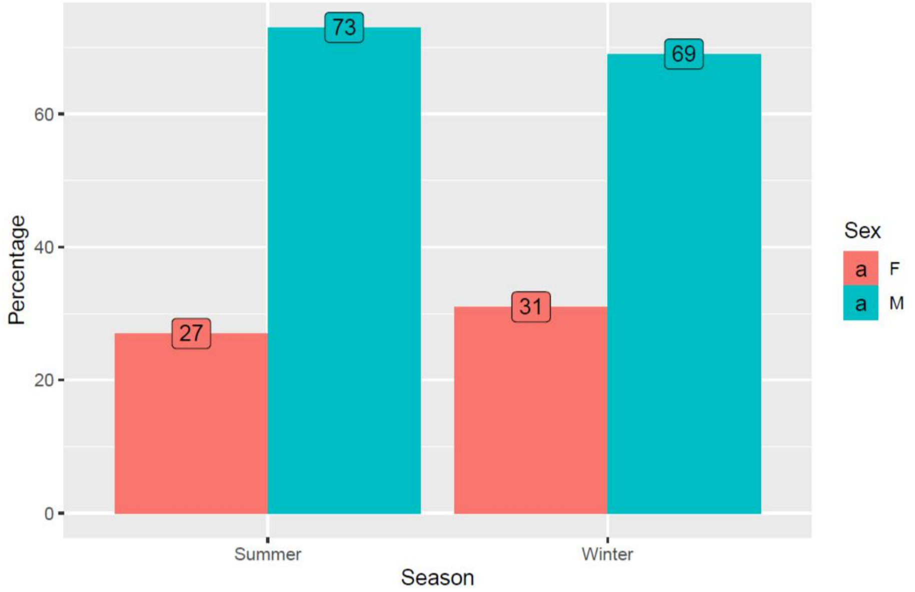
- Missing values in the medals column are considered as if the player did win any medal that year.

- Outliers will be removed from the data.

3. Exploratory Data Analysis

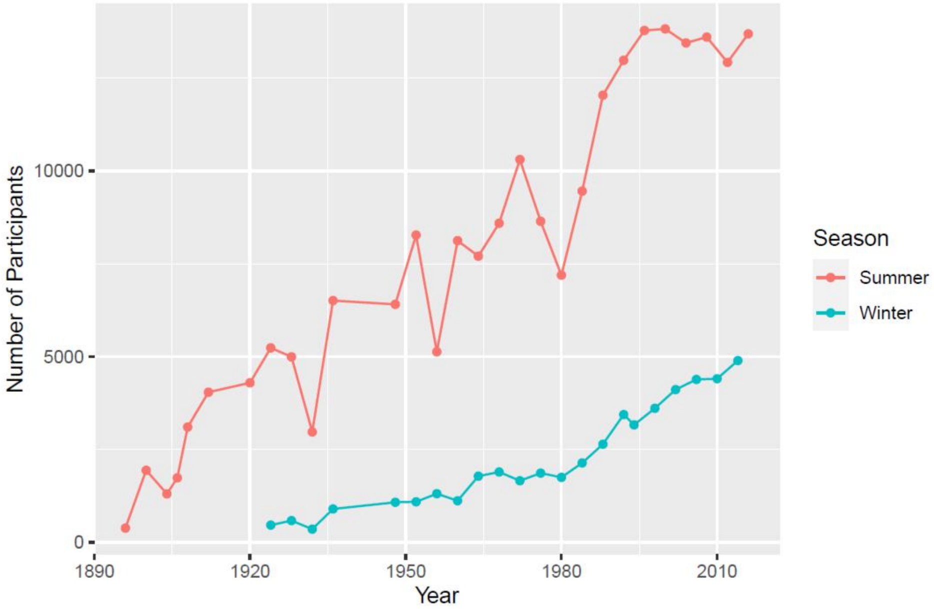
The Exploratory Data Analysis is performed to get the various trends and figures of changes and development happened in Olympics over the past 120 years.

Male vs Female Participants



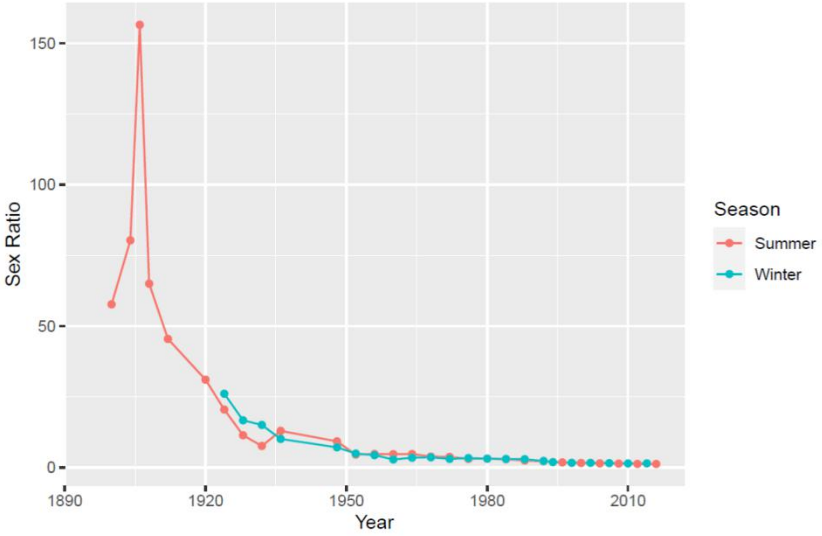
The above bar graph shows the male and female participation percentages all across the globe. During the summer season, females and males have a participation percentage of 27 and 73, respectively, while in the winter season, the percentages are 31 and 69. The percentage of women's participation during summer and winter is almost the same.

Evolution of Male vs Female participation overtime



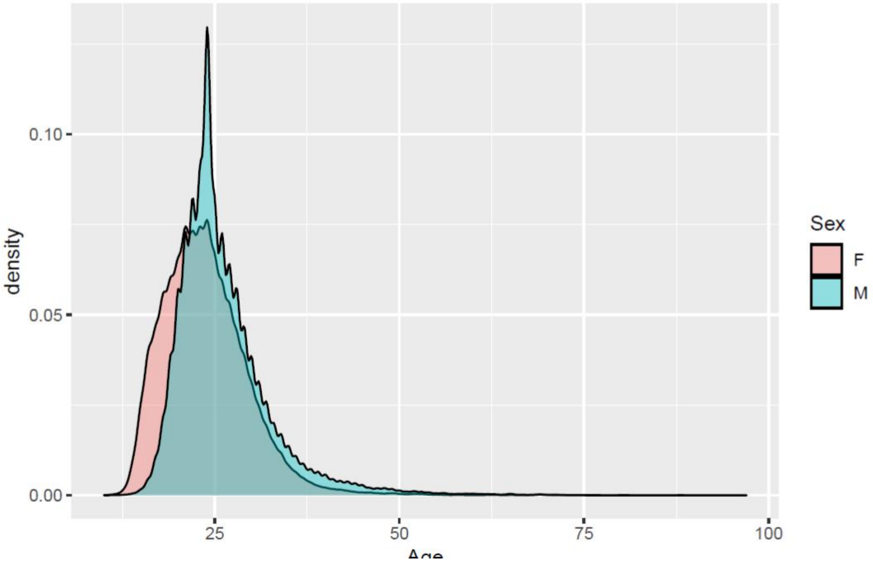
The above line graph shows the participation of players from all over the world over the years. We can see there is a steep increase in both the summer and winter seasons, but the summer's growth is huge. From this, we can conclude that many countries prefer the participation of their players during the summer season over the winter season.

Trend of Sex Ratio over the years



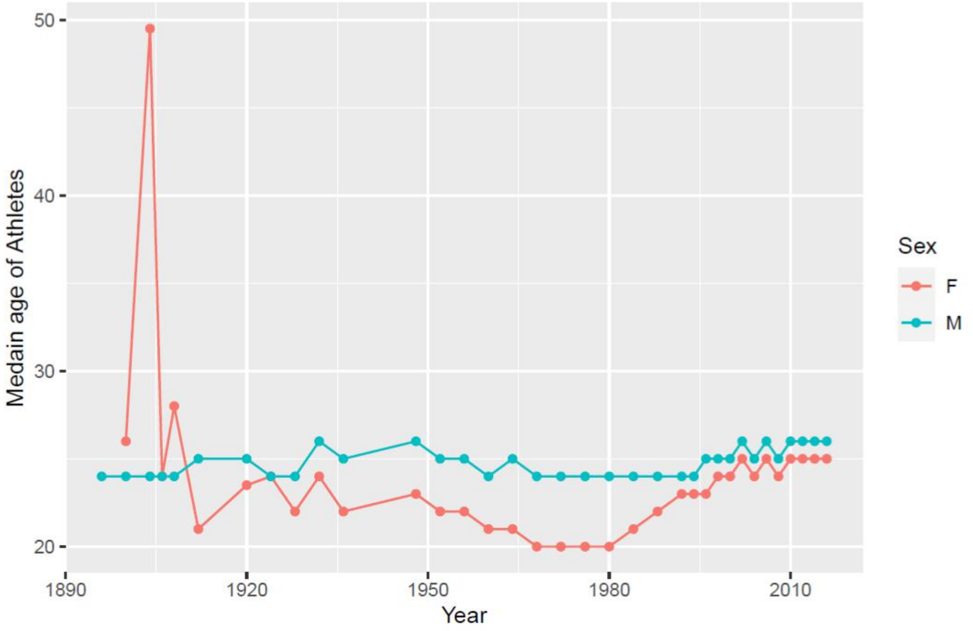
The above line graph shows the trend of the sex ratio in the Olympics from 1890 to 2010. There was a peak in the 1900s due to little or no female participation in the Olympics, which explains the peak. But as the years passed, the ratio was equal to that of 1950. There were a lot of women who competed in the Olympics.

Distribution of Age



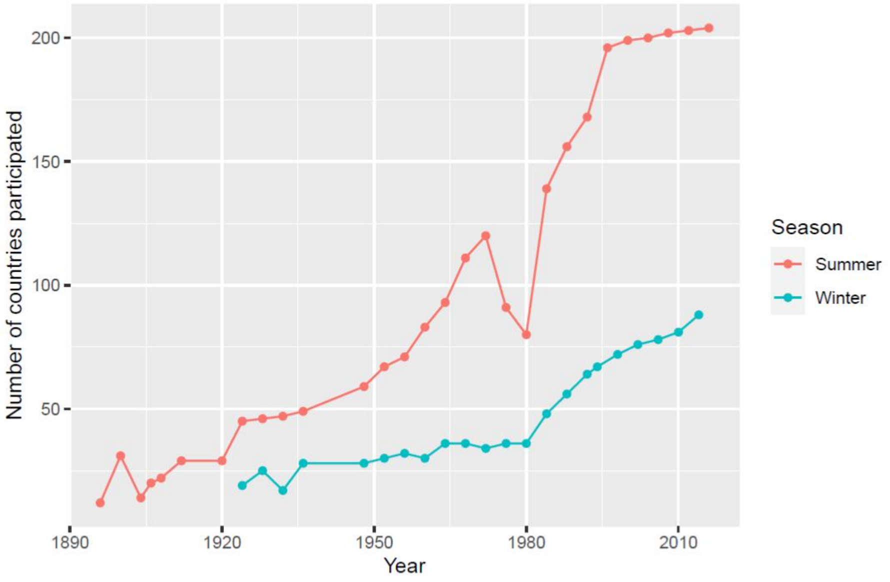
The above area graphs depict the distribution of age by sex in the Olympics. The distribution for both males and females is almost uniform, with peaks at the age of 25. We can see that the peak for the male distribution is high when compared to the female distribution, as the population of male players is significantly greater when compared to female players at the age of 25.

Median age of Male and Female players over the years



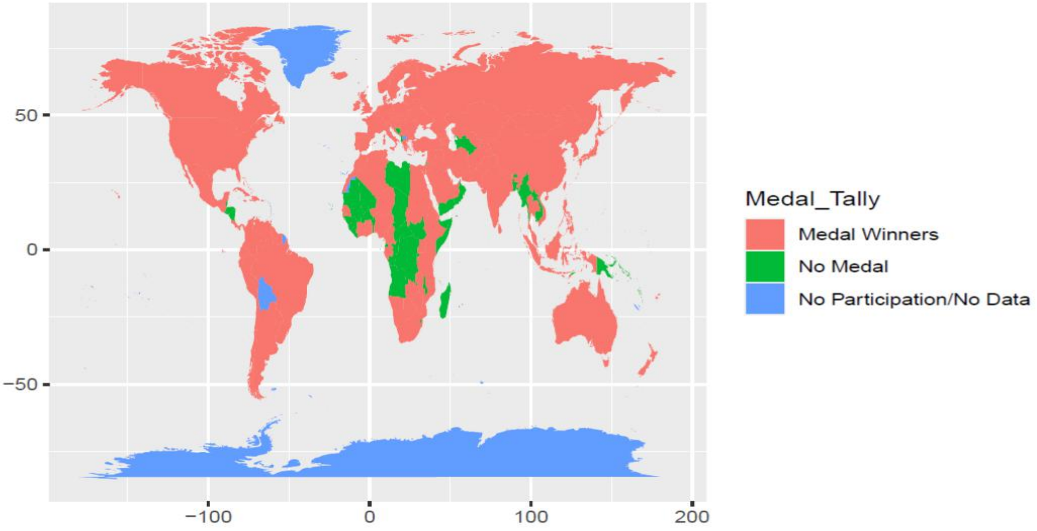
The above line graph conveys the trend of the median age of male and female players over the years in the Olympics. Initially, the median age for women was 50, which is quite interesting, and as the years passed by, the median age went down and normalized within the range of 25 to 30.

Trend of participation of countries in the Olympics



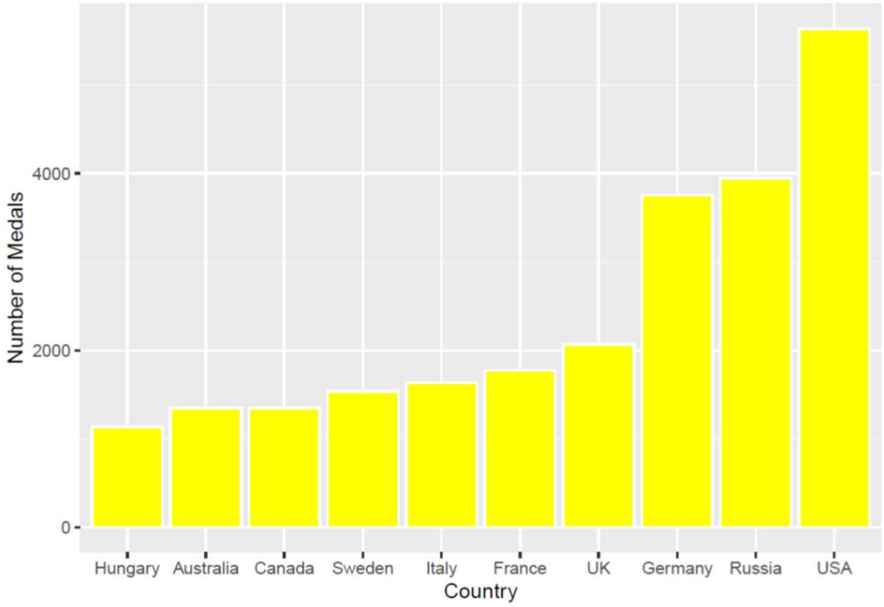
The above line graph shows the trend of participation of countries in the Olympics over the years in both seasons. We can see that there has been an increase in the participation of countries in the Olympics over the years, but the increase in the summer season is significant when compared with the winter season. From this, we can conclude that many countries often opt for the summer season over the winter season.

Distribution of medals won over the world



On the above colour-coded world map, we can see the different regions that won medals, who won no medals, and who had no participation. Even if a region wins one medal, that is counted as a region that won one medal. From the above map, we can see that there are significant regions in South Africa that did not win any medals. This could be due to the poor economy of the country.

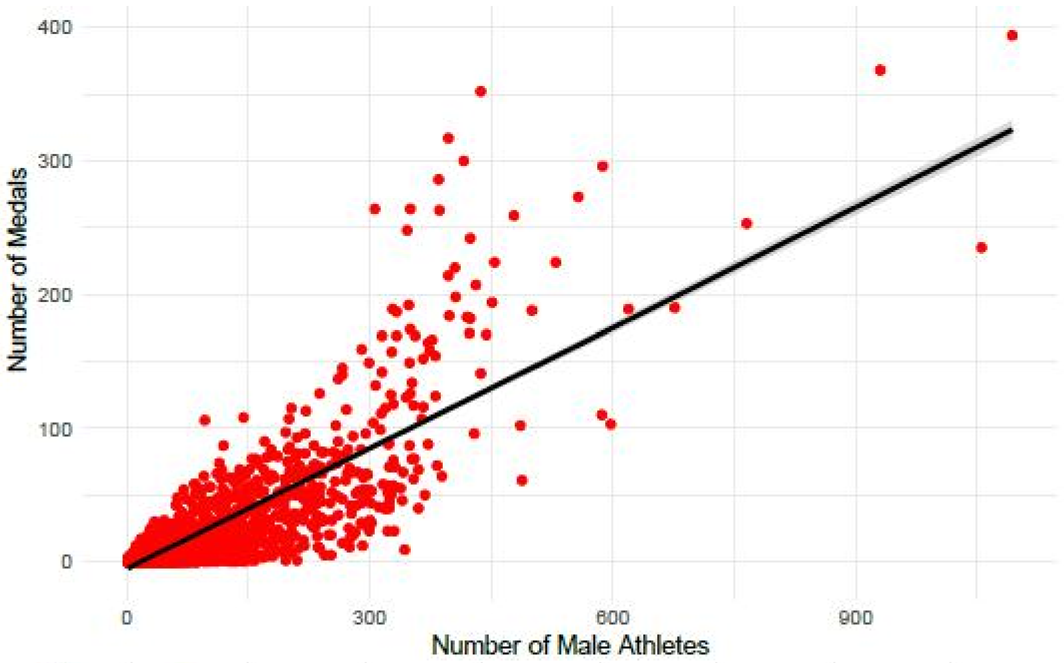
Top 10 countries with highest number of total Medals won

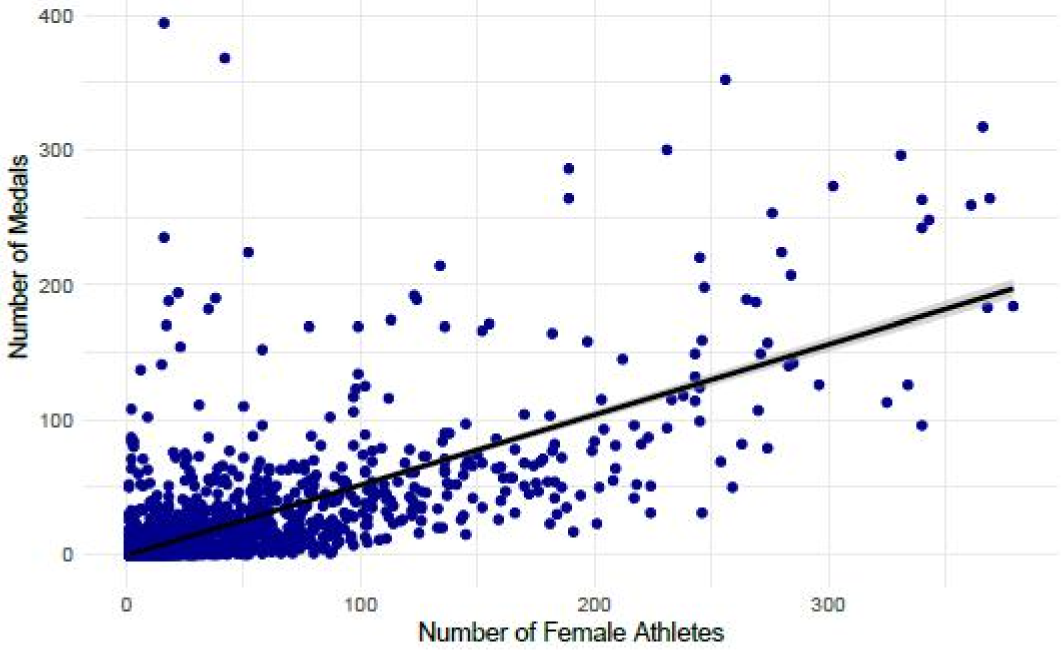


The above bar graph compares the total medals won by each country with the others. This shows only the top 10 countries who secured the highest number of medals over the years. We can clearly see that the USA has secured the maximum number of medals in the Olympics, followed by Russia, Germany, UK, France, etc.

4. Understanding the impact of factors on medal won

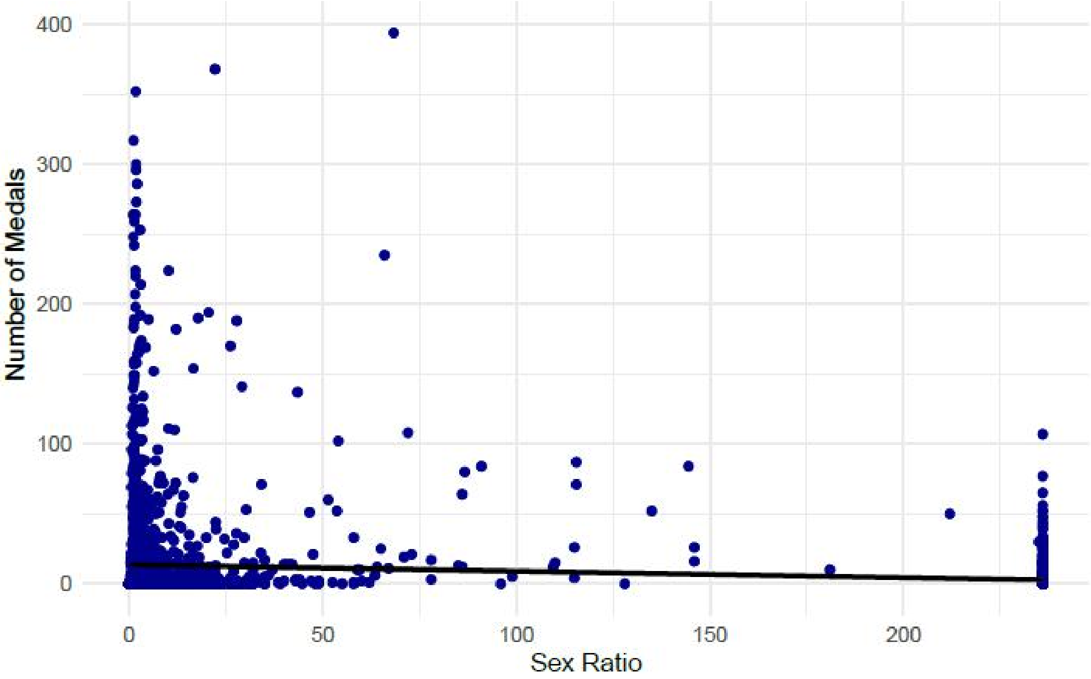
Number of participants





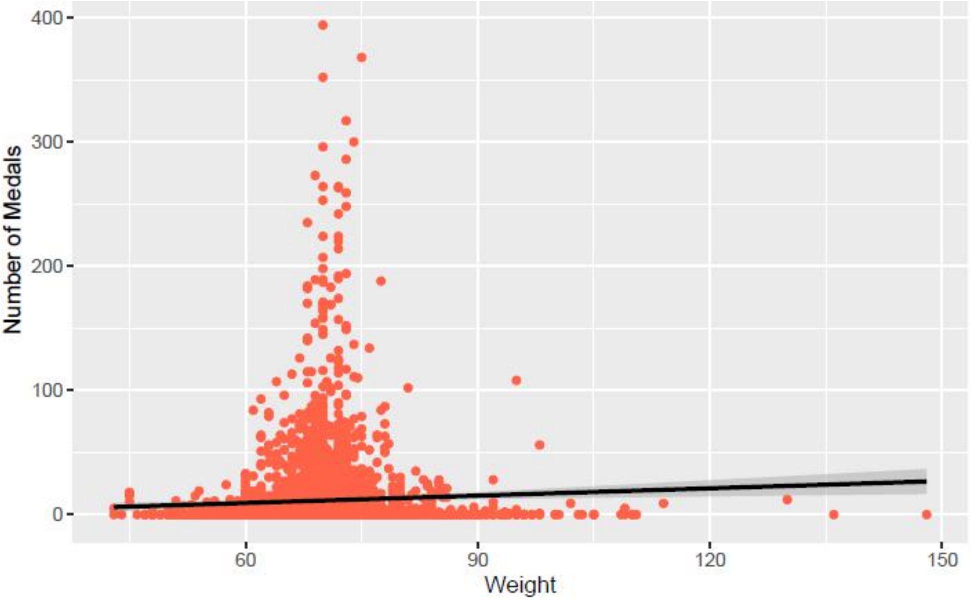
From the above scatter plots, we can clearly conclude that as the number of participants increases, the number of medals won by a country also increases. The same is the case for males and females. But the number of male participants is slightly more correlated to the number of medals won than the number of female participants. To gain a better understanding, we examined the relationship between the sex ratio and the number of medals won.

Sex Ratio



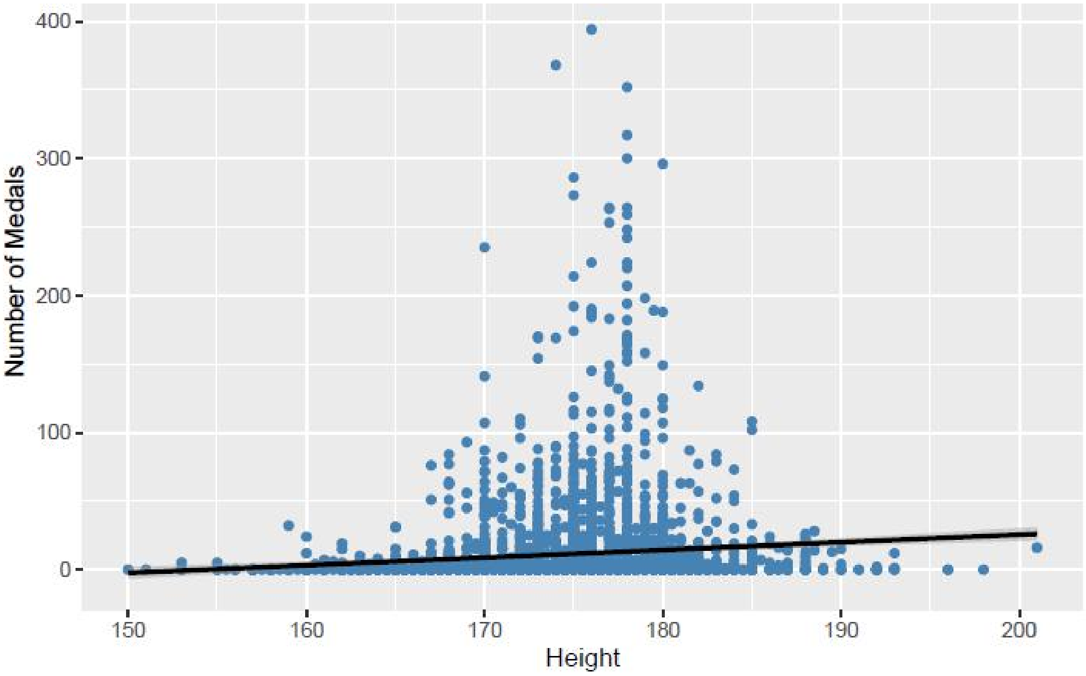
An equal number of male and female participants has no impact on the number of medals won. But a sufficient number of male and female participants from a country is needed. A country with only male participants will lose the opportunity to participate in sports events meant for females. This will definitely have a huge impact on the total number of medals won.

Effect of Weight



Weight doesn't have much impact on the number of medals won.

Effect of Height



Height doesn't have much impact on the number of medals won.

End Notes

From the above analysis and charts, we can trace the evolution that has happened in the Olympics for the past 120 years. We can clearly understand that as the number of players increases, the number of medals won by a country will also increase. Other physical factors, like height, weight, age, etc., don't have much impact on the number of medals won. Each sport has its own needs, and the Olympics is a collection of multiple sport events. So physical factors won't have much impact on the total number of medals won in the Olympics. Medals won depend on performance and on the hard work of the players. As these cannot be measured, it is difficult to conclude which country will top the table. Some predictions can be made from the previous year's trends, but confirmed results can never be given.