# Olympic Data Analysis

Swarnajit Podder, Pratyusha Sarkar, Chirag Garg, J S K L N M Subbarayadu

Data Science Lab 1 Project Group 5

## Introduction

Olympics is considered as one of the most prime platforms for the players across the countries all over the world to showcase their talents. Athletes becomes pride of their countries, more importantly, being icon of great inspiration to the young generation.

The first ancient Olympics games were held in Olympia, Greece at 776 BC as a part of some religious festival and it continued till 393 AD in every four years at different cities of Greece. After almost 1500 years it reappeared at 1896, which was the first modern Olympic games held at Athens, Greece. Since then every four years, the biggest festival of sports is celebrated at different corners of the world. Thus being enriched with glorious history, it becomes more than some multi-sport championship. For years, people are trying to scrutinize the games under the lens to understand global history including geopolitical dynamics, evolution of human society, women empowerment etc.

## Motivation and Proposed Approach

Various scenarios comes to our mind when we look into evolution of Olympic Games over the years like increase in participation of countries, increase/decrease in number of events, improvement of performances, participation ratio of men and women. In this project, the target is to analyze that history of olympics to find possible answers of how different factors contributes to the result of the games over years. To determine these factors and perform a comparative study on these factors, following steps are followed. We try to show them in a flowchart.

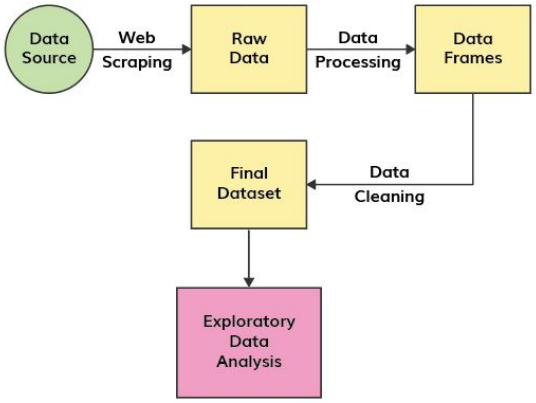



Figure 1: Flowchart of work

## Data Source

In order to perform analysis we require a large amount of data on which we can apply various techniques to reach to a particular conclusion. Earlier we wanted to scrape the data from the official website of Olympics, but it created some issue by blocking the user to access their website html for scraping. So we have used another website. This website consists the information about the players and their entire details like their Gender,Height, Weight, Country for which they play, Medals won (Gold, Silver and Bronze) and many more. These data are used analyze the performance of players over the year and for each event of one particular year. We have collected data only on last 10 summer Olympics, i.e. from Los Angeles 1984 to Tokyo 2020. Also we collected data on GDP of every countries from this website. This dataset consists each country code,and GDP of that particular country over the years.

## Data Scraping

**Required Libraries**
To scrape the data from olympedia.org some libraries are required such as :

- rvest
- tidyverse
- eeptools
- lubridate
- stringi

**Web Scraping : Steps**

**Scraping for list of medal winners**

- First load the website olympedia.org in R.
- Select years from 1984 to 2020 except three youth Olympics such as Singapore 2010, Nanjing 2014, Buenos Aires 2018.
- Create the url link based on the unique ID of each Olympics, like this :

```
## [1] "http://www.olympedia.org/editions/21/medal"
## [2] "http://www.olympedia.org/editions/22/medal"
## [3] "http://www.olympedia.org/editions/23/medal"
## [4] "http://www.olympedia.org/editions/24/medal"
## [5] "http://www.olympedia.org/editions/25/medal"
## [6] "http://www.olympedia.org/editions/26/medal"
```

- Now we find the **country codes** and **country names** of those countries which won the medals at least once.
- A function was created to accumulate the details of each of the olympic, one at a time.

Using `html_table()` we got a table like this.

```
## # A tibble: 6 x 7
##   X1                          X2                  X3      X4    X5    X6    X7
##   <chr>                       <chr>               <chr>   <chr> <chr> <chr> <chr>
## 1 Sport/Event                 Gold                ""      Silv~ ""    Bron~ <NA>
## 2 Archery                     Archery             "Arche~ Arch~ "Arc~ Arch~ Arch~
## 3 Individual, Men             Darrell Pace        "USA"   Rick~ "USA" Hiro~ JPN
## 4 Individual, Women           Seo Hyang-Sun       "KOR"   Li L~ "CHN" Kim ~ KOR
## 5 Artistic Gymnastics         Artistic Gymnastics "Artis~ Arti~ "Art~ Arti~ Arti~
## 6 Individual All-Around, Men Koji Gushiken        "JPN"   Pete~ "USA" Li N~ CHN
```

- Equating the **first two columns of the table** we can find the name of the main events.Repeat each event name suitable number of times to arrange the dataframe properly. After that we make a separate column for subevents.
- Now to identify the type of each games, in the medal winners list in the initial dataframe, if there is a country name then that event is a *team event* otherwise it is an *individual event*. One column is added to it to list down the type of medals (Gold, Silver, Bronze) won by the players.

```
##      Events        Sub.Events       Type  medal
## 1   Archery   Individual, Men individual   gold
## 1.1 Archery   Individual, Men individual silver
## 1.2 Archery   Individual, Men individual bronze
## 2   Archery Individual, Women individual   gold
## 2.1 Archery Individual, Women individual silver
## 2.2 Archery Individual, Women individual bronze
```

- Name of individual players and their countries are listed down from the initial table. On the other hand, for the team events, put **NA** in corresponding rows.
- Here, a big issue occurred in case of presence of *joint winners* in an event. If there are joint winners, the names and their country codes have been joined together in the inital table. Since, country codes are always of 3 letters, they could be easily handled by separating each group of 3 letters. To seperate the names of the players, a pattern was followed initially - *a small alphabet followed by a capital alphabet.* But there were several exceptions of this pattern, like -
  - In many cases, names contain some alphabets of their own languages. These letters were listed seperately and another condition was added to the pattern to take care of them.
  - In some cases, names itself contained a small letter followed by a capital letter (e.g. *Julianne McNamara*). These cases had to handle manually.
- After seperating the names and countries of the joint winners they had to be appended in proper position of rows in the dataframe.

Thus names and countries of all medal winners were listed .

```
##    Events        Sub.Events       Type  medal       html_medal
## 1 Archery   Individual, Men individual   gold     Darrell Pace
## 2 Archery   Individual, Men individual silver    Rick McKinney
## 3 Archery   Individual, Men individual bronze Hiroshi Yamamoto
## 4 Archery Individual, Women individual   gold    Seo Hyang-Sun
## 5 Archery Individual, Women individual silver      Li Lingjuan
## 6 Archery Individual, Women individual bronze       Kim Jin-Ho
##   html_country_code
```

```
## 1               USA
## 2               USA
## 3               JPN
## 4               KOR
## 5               CHN
## 6               KOR
```

**Scraping for details of individual players**

- In the website, the players' name were hyperlinked with the pages containing their personal details and history of medal winnings.
  - **Age Calculation :** In each pages, Date of birth are given for each players. Using that, age of the players were calculated in the year of the corresponding Olympics. Some discrepencies had to be sorted during this, like :
    - Some DOB was along with their birth places, from which dates were seperated.
    - Some DOB didn't have the day, only months and years were available. We considered the day 1 of the corresponding months in these cases.
  - **Height and Weight :** Heights and weights of the players was given in a single line under 'Measurements'. Those had to be separated and units had to be removed to get the numerical values of the measurements.

Following all of the above steps, 10 dataframes were created, one for each Olympics. Here is a glimpse of one of the final tables.

```
##   year  Events         Subevents       Type  Medal           Player Country Age
## 1 1984 Archery   Individual, Men individual   gold     Darrell Pace     USA  28
## 2 1984 Archery   Individual, Men individual silver    Rick McKinney     USA  31
## 3 1984 Archery   Individual, Men individual bronze Hiroshi Yamamoto     JPN  22
## 4 1984 Archery Individual, Women individual   gold    Seo Hyang-Sun     KOR  17
## 5 1984 Archery Individual, Women individual silver      Li Lingjuan     CHN  19
## 6 1984 Archery Individual, Women individual bronze       Kim Jin-Ho     KOR  23
##   Gender Height Weight
## 1   Male    180     64
## 2   Male    170     59
## 3   Male    170     71
## 4 Female    171     67
## 5 Female    166     62
## 6 Female    164     54
```

**Scraping for GDP Data**

- First go to the GDP website From this website, the country codes were screpd and using those go to the website of each of them.
- Data on GDP was collected only for the years of last 10 summer Olympics (1894 - 2020).
- Finally, from the previous data on medal winners, data on numbers of medals won by each countries are collected and attached to this dataframe on GDP.
  The data looks like this :

```
##   Year Country Medals Country_Name   Longitude
## 1 1984     ALG      2      Algeria    1.659626
## 2 1984     AUS     24    Australia  133.775136
## 3 1984     AUT      3      Austria   14.550072
## 4 1984     BEL      4      Belgium    4.469936
## 5 1984     BRA      8       Brazil  -51.925280
## 6 1984     CAN     44       Canada -106.346771
```

## Some Key Questions

The purpose of our analysis is to answer some, if not all, of the interesting questions like this :

1. Which countries dominate in which Olympic Sports?
2. Which factors affect the total number of medals won by different countries?
3. Does the Olympic games show gender equality?
4. How does age of a player impact winning a medal in Olympics?
5. What is the effect of Height and Weight to win a medal in different games?
6. How is GDP of a country related to number of medals won by a country?
7. How have number of medals won by a country varied over the years?

## Visualizations and Analysis

To find the possible answers of the questions, we are using Exploratory Data Analysis technique. Exploratory Data Analysis (EDA) is *an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods* (**Source : Wikipedia**). With the help of EDA we try to analyse the questions visually apart from applying any statistical tools. Varoius types of plots are used here to explain things. Some of them are -

- Scatter Plot
- Bar Diagram
- Density plot
- Box Plot
- Dot Chart
- Plot on world map etc.

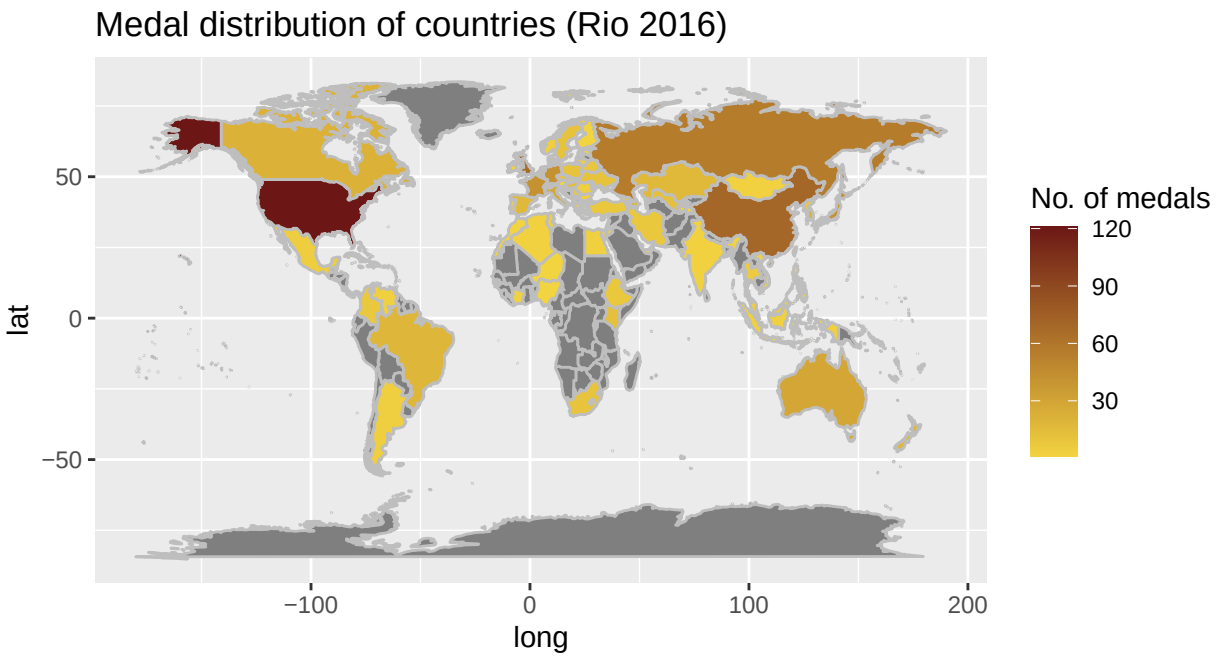
### Plot of Medal Distribution over Years

### Plot on World Map

Our target is to plot the number of medals won by different countries over years using colors according to medal distribution. To do this first the dataset 'world' is loaded and a column is added to it which contains number of medals won by different countries in a particular year. Here is a glimpse of the dataset.

```
##            long      lat group order region subregion map_medal
## 12463 -53.08228 2.201709   238 12463 Brazil      <NA>        19
## 12464 -53.00972 2.181739   238 12464 Brazil      <NA>        19
## 12465 -52.96484 2.183545   238 12465 Brazil      <NA>        19
## 12466 -52.90347 2.211524   238 12466 Brazil      <NA>        19
## 12467 -52.87041 2.266650   238 12467 Brazil      <NA>        19
```

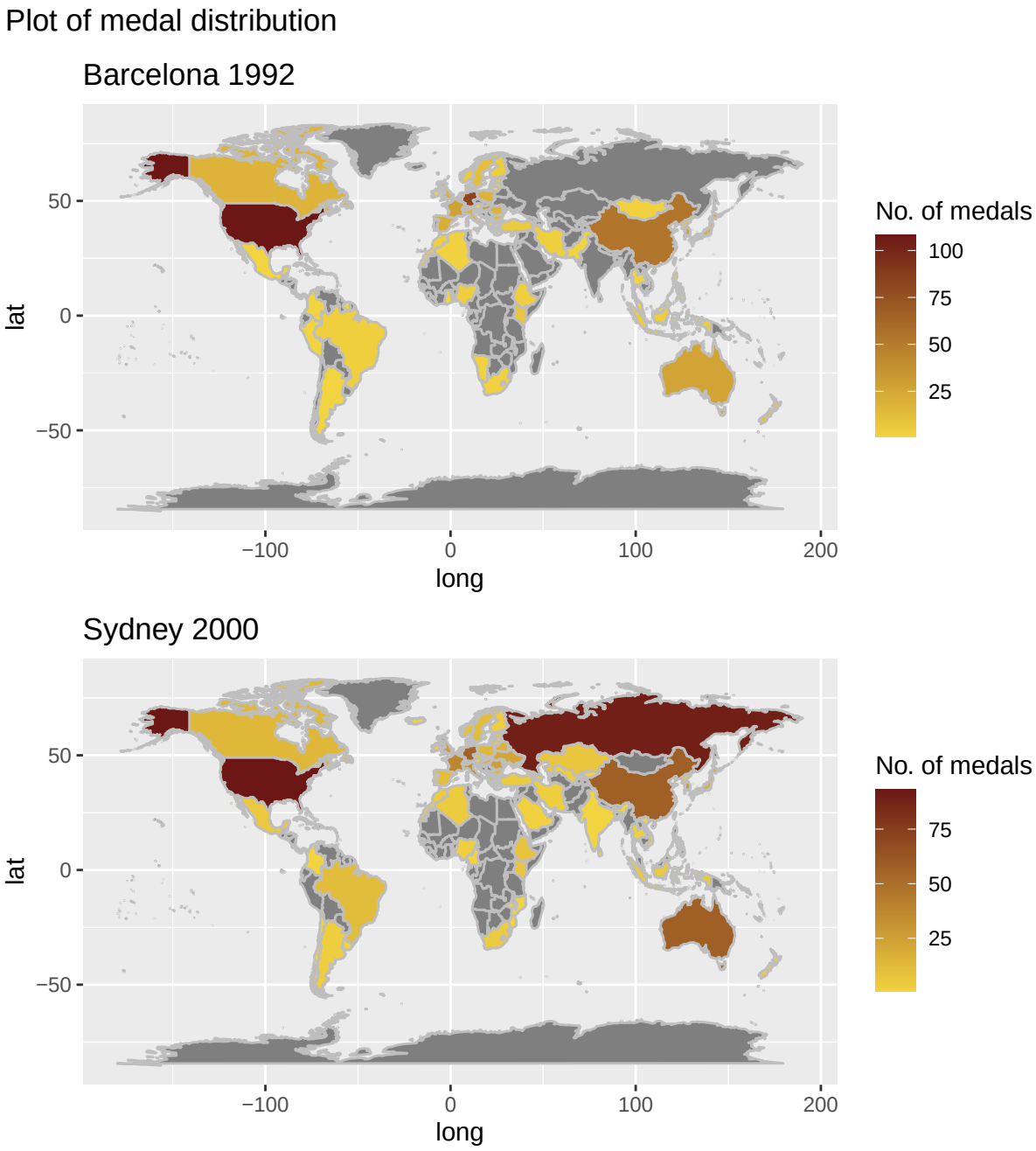
Now to plot the world map based on the data, we need **geom_polygon** function from *ggplot2* package. Then on the map, colors are given according to number of medals on by different countries in that particular year. We get a plot as follows.




**An Observation : Host Country Effect**

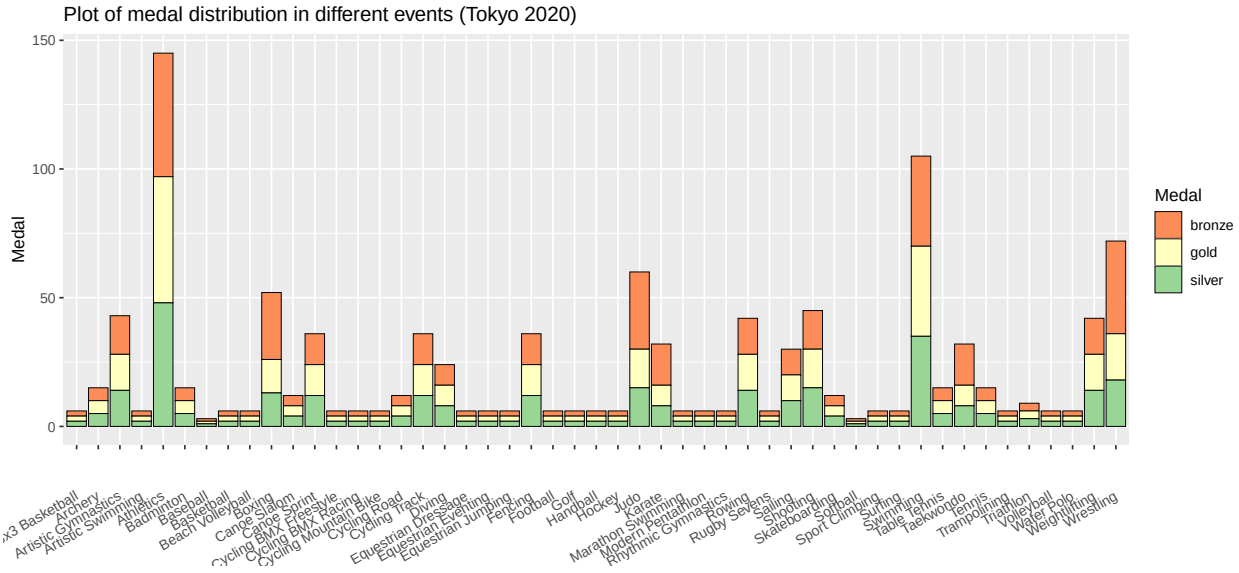
We plot the above graph for the years 1992 and 2000.

From the graph below, it is observed that the medal distribution of Australia had increased significantly from 1992 to 2000. Now, in 2000, the host country was Australia, and that is the probable reason of this drastic increase as there was a significant improvement of infrastructure and huge increase in investment to arrange the events. This is known as **Host Country Effect**.



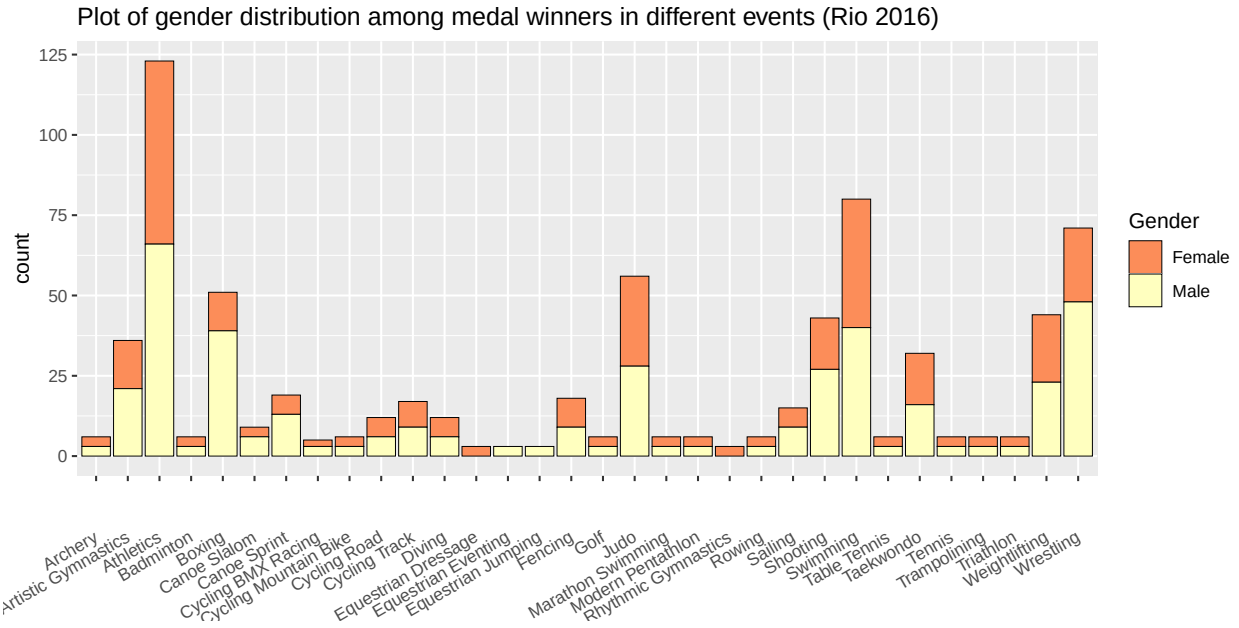

**Plot of medal distribution over different events in an Olympic**

To show number of medals won by the players in different events in a particular olympic we draw a *Stacked Bar Diagram.* In every bar, different types of medals (Gold, Silver, Bronze) are shown by different colors. Here is the plot.




### Plots on Gender Distribution

To show the gender distribution among the medal winners in an Olympic, another *Stacked Bar Diagram* was plotted for each of the Olympics. In each bar, Number of Male and Female medal winners are denoted by different colors. The plot is as follows.




### Does Olympics show gender equality?

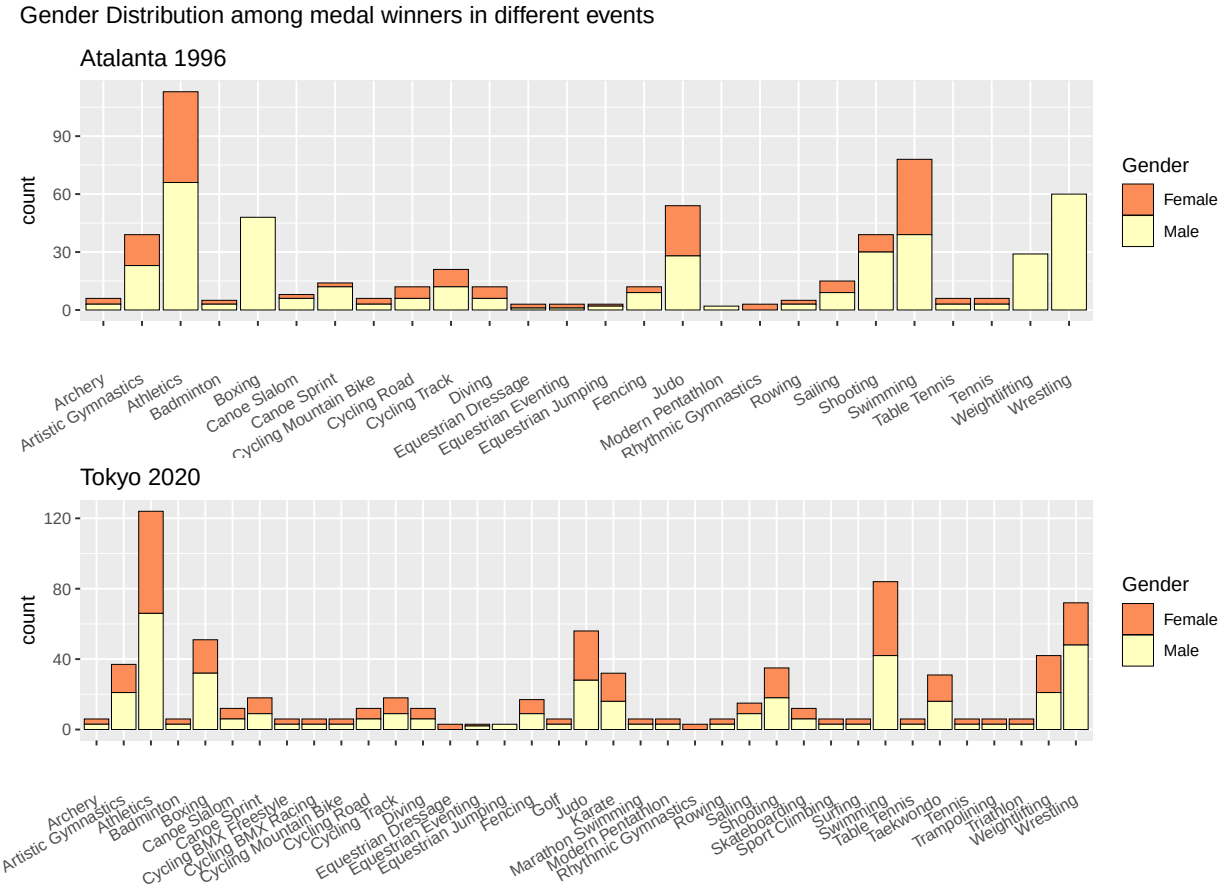
Gender inequality is a very alarming issue globally. Now, one can question whether Olympics is also affected with this. For this, we have to study the ratio of number of male and female participants of different countries in Olympics over the years.

Now, the issue with this study is that we only have data on the medal winners in different Olympics. So, it

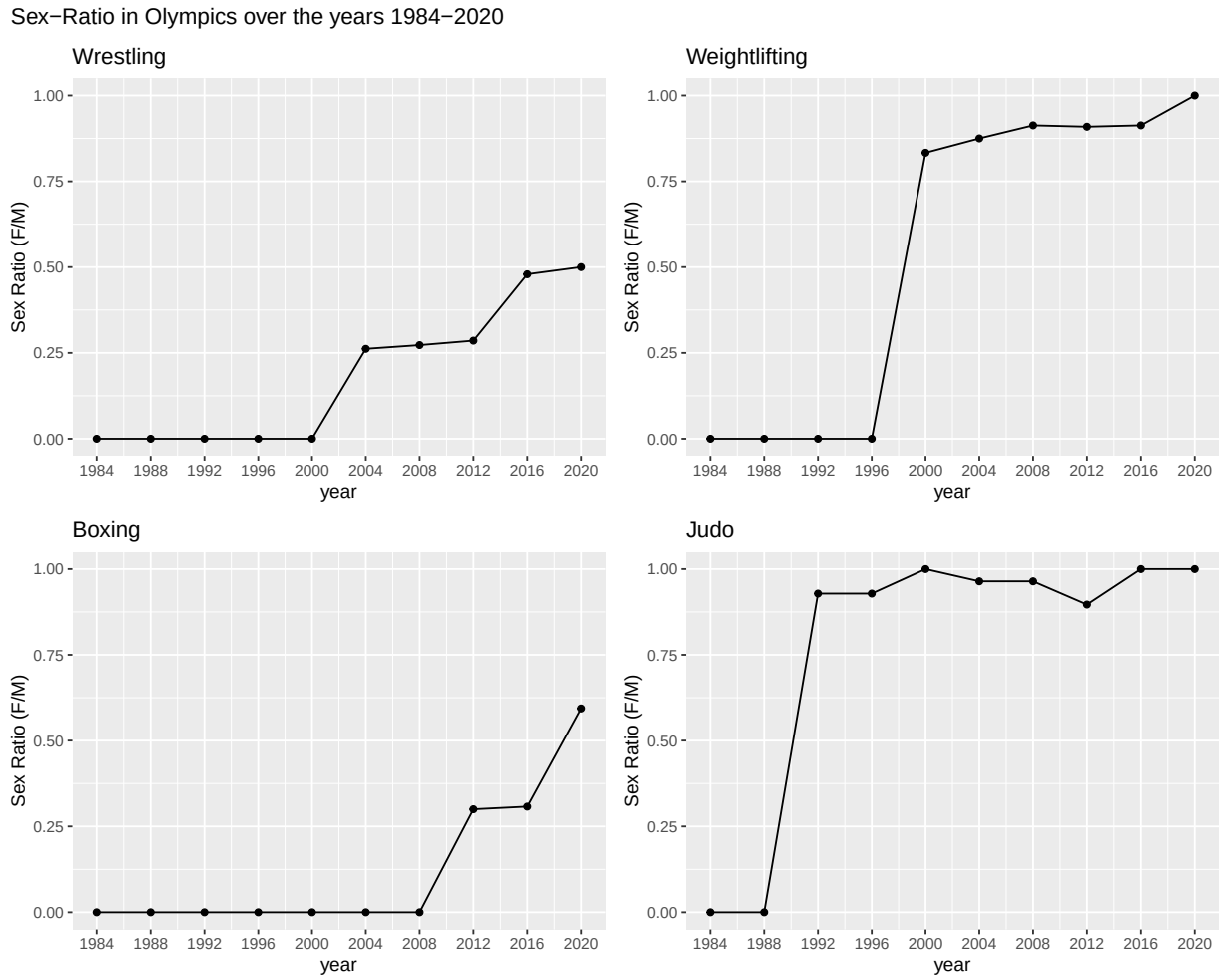
is not possible to study exact sex-ratio of different countries in the games. That's why, we tried to visualize this from a different point of View. We wanted to determine sex-ratio among the medal winners in different events in every Olympics. The data looks like this :

```
## # A tibble: 6 x 5
##    year Events               Female  Male   Ratio
##   <dbl> <chr>                 <dbl> <dbl>   <dbl>
## 1  1984 Archery                   3     3   1
## 2  1984 Artistic Gymnastics      15    24   0.625
## 3  1984 Artistic Swimming         2     0 Inf
## 4  1984 Athletics                38    67   0.567
## 5  1984 Boxing                    0    48   0
## 6  1984 Canoe Sprint              2    12   0.167
```

Since, all the events more or less have 3 medal winners, be it a male or female event, the ratio should be 1 in most of the cases. But as it can be seen from the glimpse of the data that the ratio is not so, even there is some 0's and *inf*'s (i.e. no male/female medal winners) which only implies that there is not the equality in numbers of events for males and females in different events - inequality from the hosts themselves! Let's try to see this visually as well.




From the graph, it can easily be noticed that there was no female participation in case of physical games like **Boxing, Judo, Wrestling, Weightlifting** in 1996 Olympics, whereas in 2020, females are also actively participating in these events. So, there might be a decrease to the gender inequality issue in the games in recent years. Let's try to visualize this fact as well.




So, it is quite evident from the graphs that, the events are approaching towards achieving gender equality slowly over the years. Events like Weightlifting and Judo have completely achieved (fluctuations are due to discrepencies in the data collected), at least from the organizers' side, whereas there is still inequality in the games like Wrestling and Boxing.

### Plots on GDP

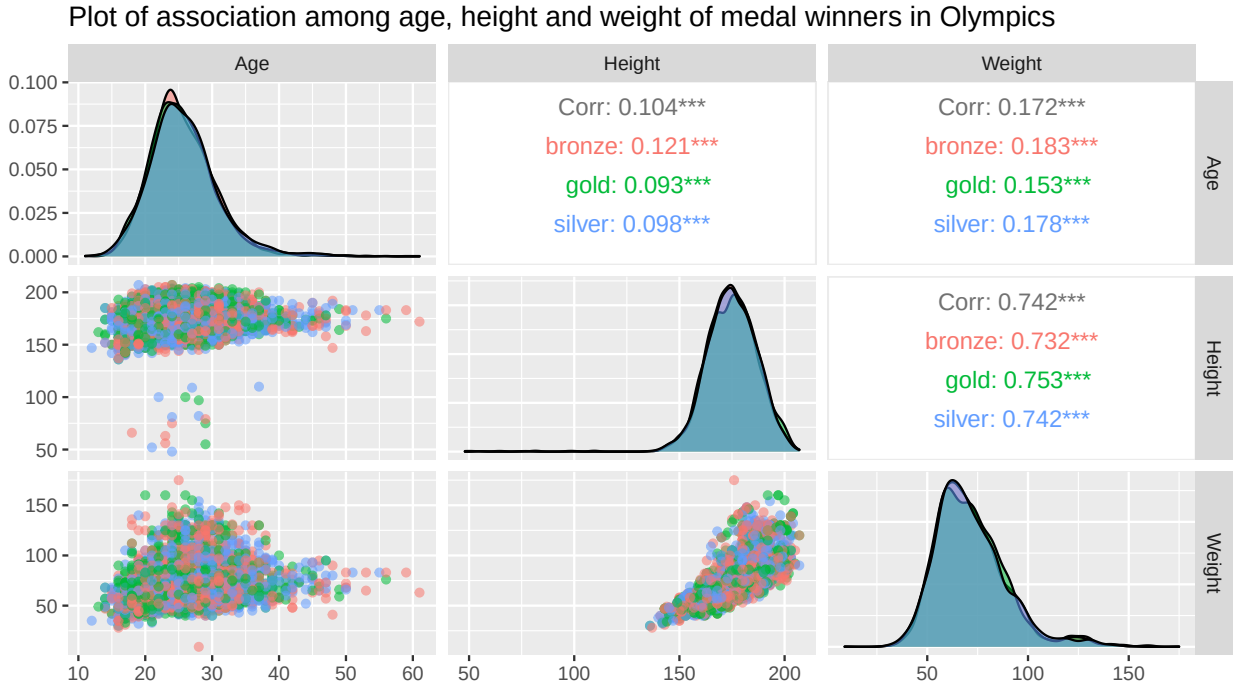
**How is number of medal won by different countries depend on GDP over the years?**
Now, if one looks into the medal tally of different Olympics over the years, one can only see the first-world countries like USA, UK, China always dominates the table. So, it is an interesting question whether GDP of a country somehow affects the medal count.

To analyse this, scatterplots are drawn on number of medals won by a country with the corresponding GDP for different years.


From the graphs, one can easily observe a positive trend in the plot. So, without fitting any statistical model, we might conclude that there is a positive correlation between GDP and number of medals won by a country in every Olympic.

### Plot on Age, Height and Weight

**How does age, height and weight of a player affect the number of medals won?**
To analyse the effect of age, height and weight, first we want to analyse the association between each of the factors. To do this, we use **ggpairs** function from `GGally` package to obtain a plot like this :

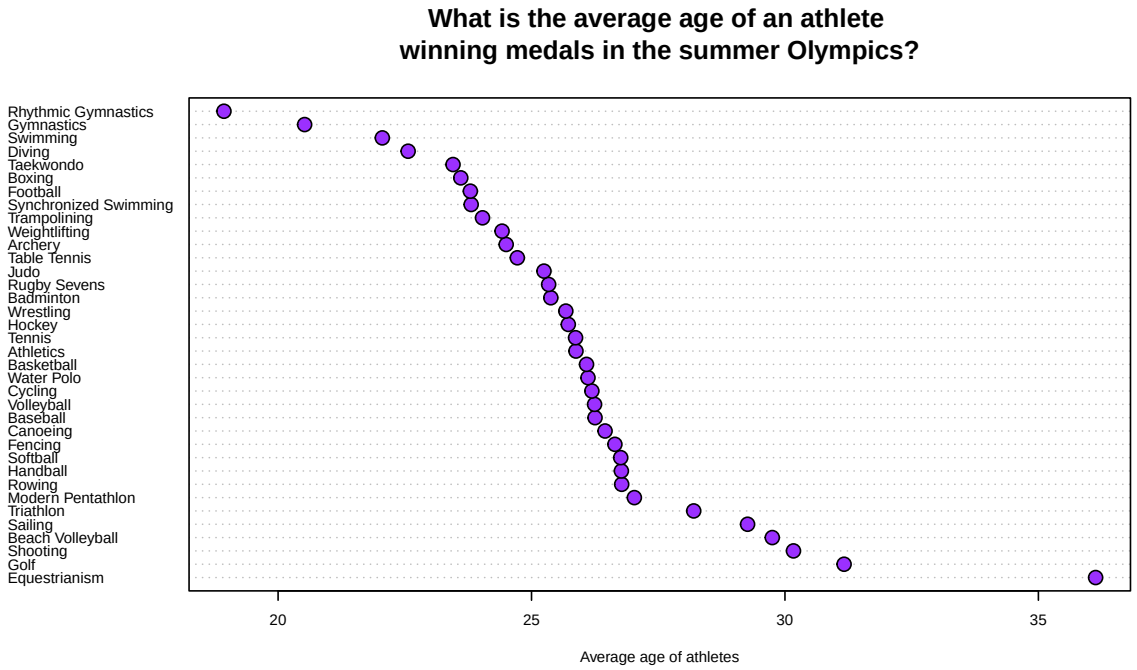



We can see that, there isn't much association of age with the other two but height and weight of players are **highly positively correlated** among themselves.

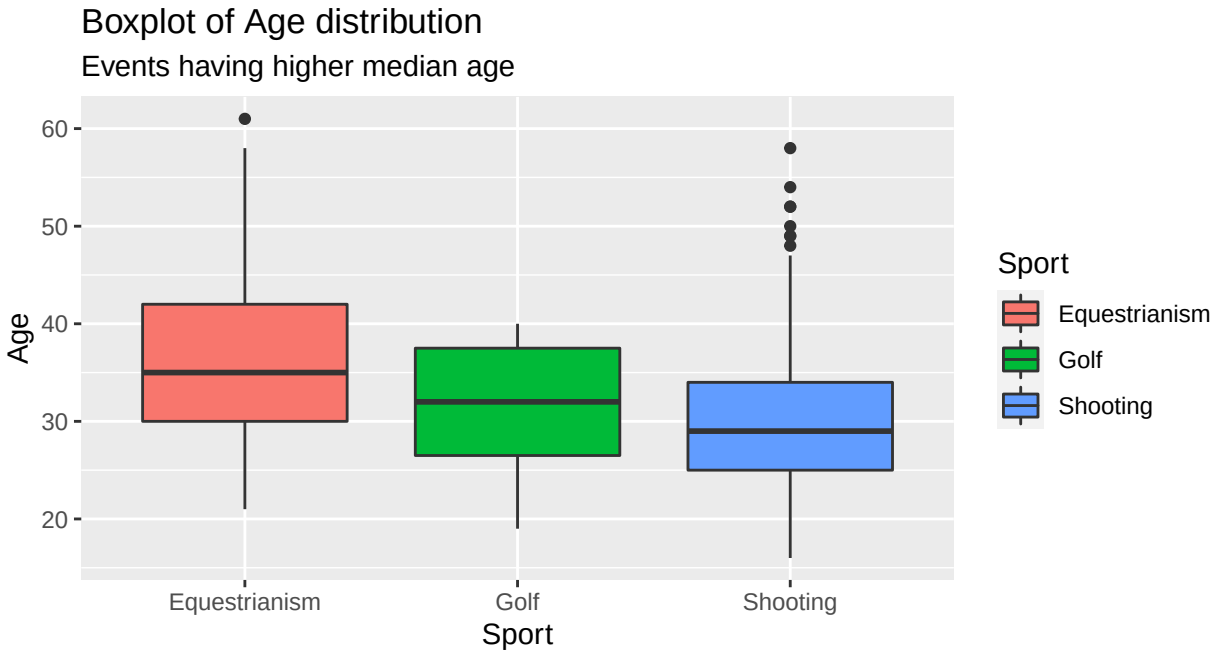
Now, it is more interesting to understand the effect of these factors in different events seperately.

### Distribution of Age in different events

Now we draw a dotchart of ages of the medal winners for different events. The graph looks like this :

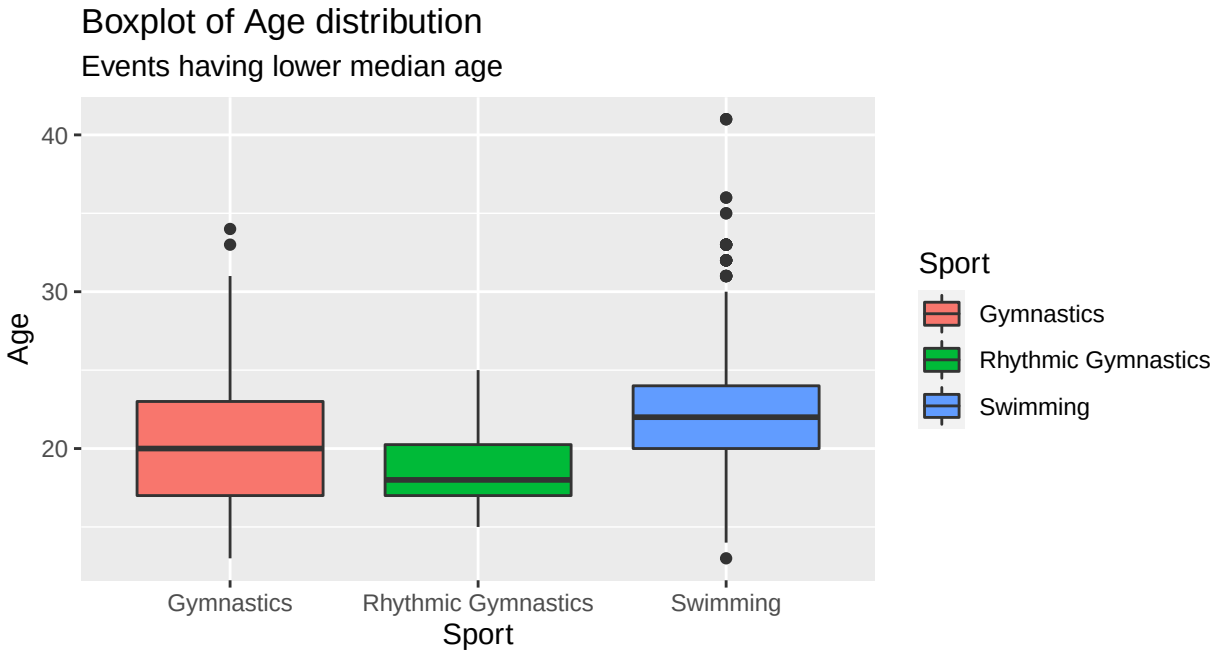


From the graph, it can easily been observed that, on and average age of medal winners varies in between 20 and 35 years, more or less. In particular, sports like **Equestrianism, Shooting and Golf** comparatively aged players are winning the medals whereas sports like **Rhythmic Gymnastics or Gymnastics** are dominated by the teenagers or players in early 20s. Let's study the variations in age in these events using *Boxplot*.




From the plot, observe that both teenagers and players having age over 40 are winning medals in these events. So, these events have a wide range of participations in terms of age groups.

On the other hand, observe this boxplot :



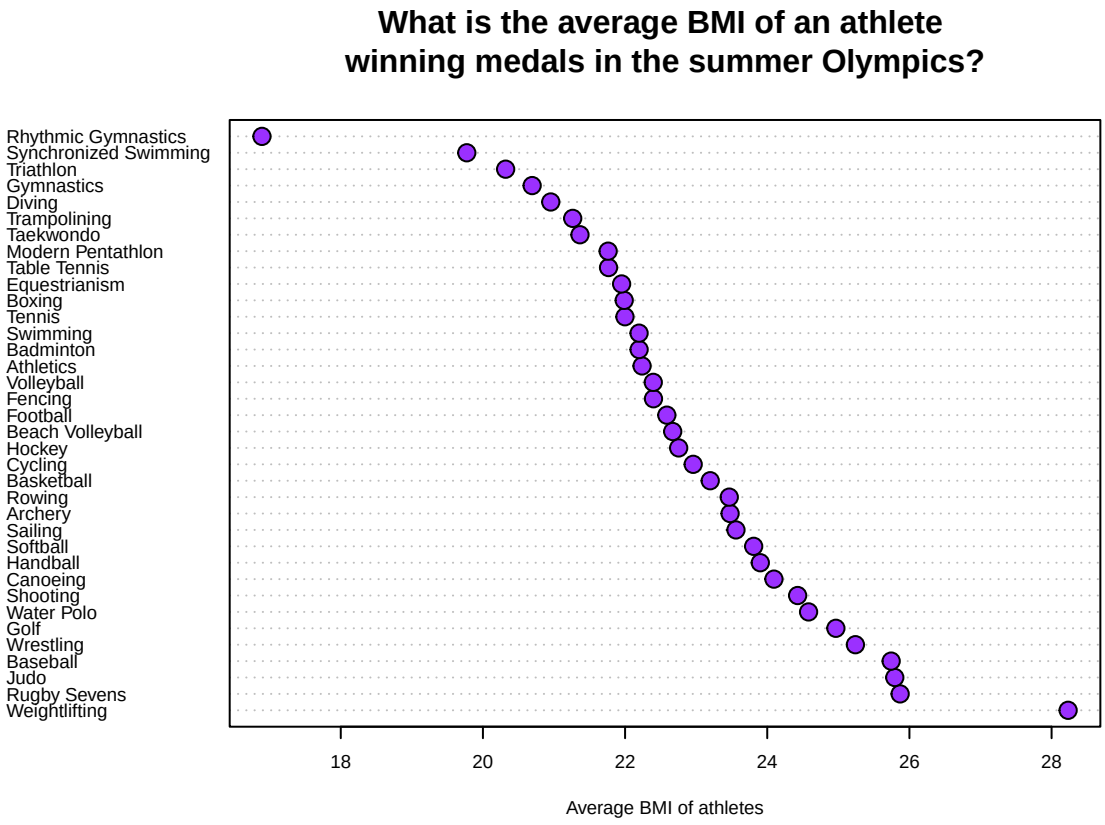

Here, the median age is very low (around 20) but the variations are quite high for Gymnastics and Swimming, some players of age more than 40 are also winning medals.

**Distribution of Height and Weight in different events**

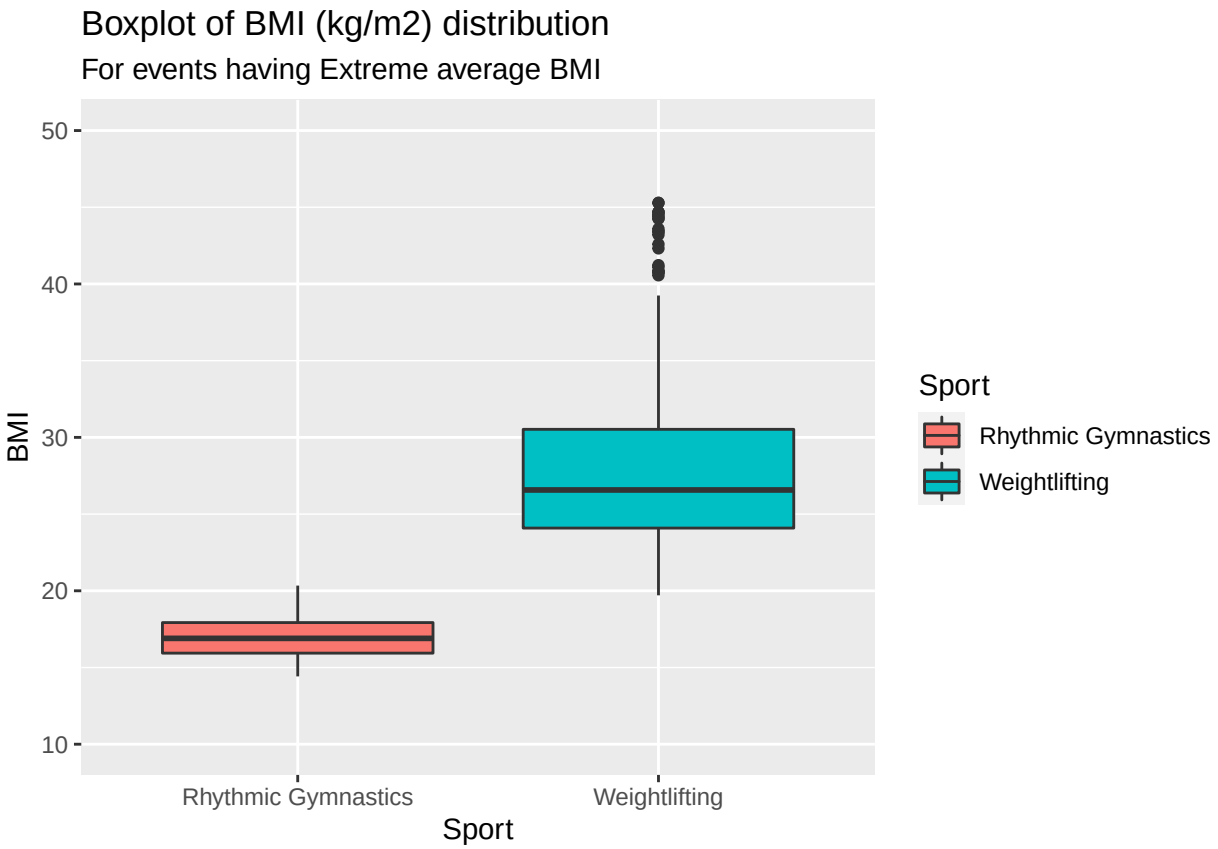
To analyse effect of height and weight of players on winning medals, first we calculate their **Body Mass Index (BMI)** using the formula :

$$BMI = \frac{Weight}{Height^2}$$

where weight is measured in kg and height in m. We draw a dotchart of BMI of the medal winners for different events. The graph looks like this :




From the graph, it is quite evident that on an average, Weightling required highest BMI to win a medal whereas for Rhythmic Gymnastics, BMI is the lowest for the medal winners. Let's see the distribution of these two events using *Boxplot*.




From the above plot, it is clear that Rhythmic Gymnastics have a very less variation of BMI among the medal winners and the average BMI is also extremely low. On the other hand, weightlifting requires very high BMI. The outliers indicate that some medal winners had extremely high BMI.

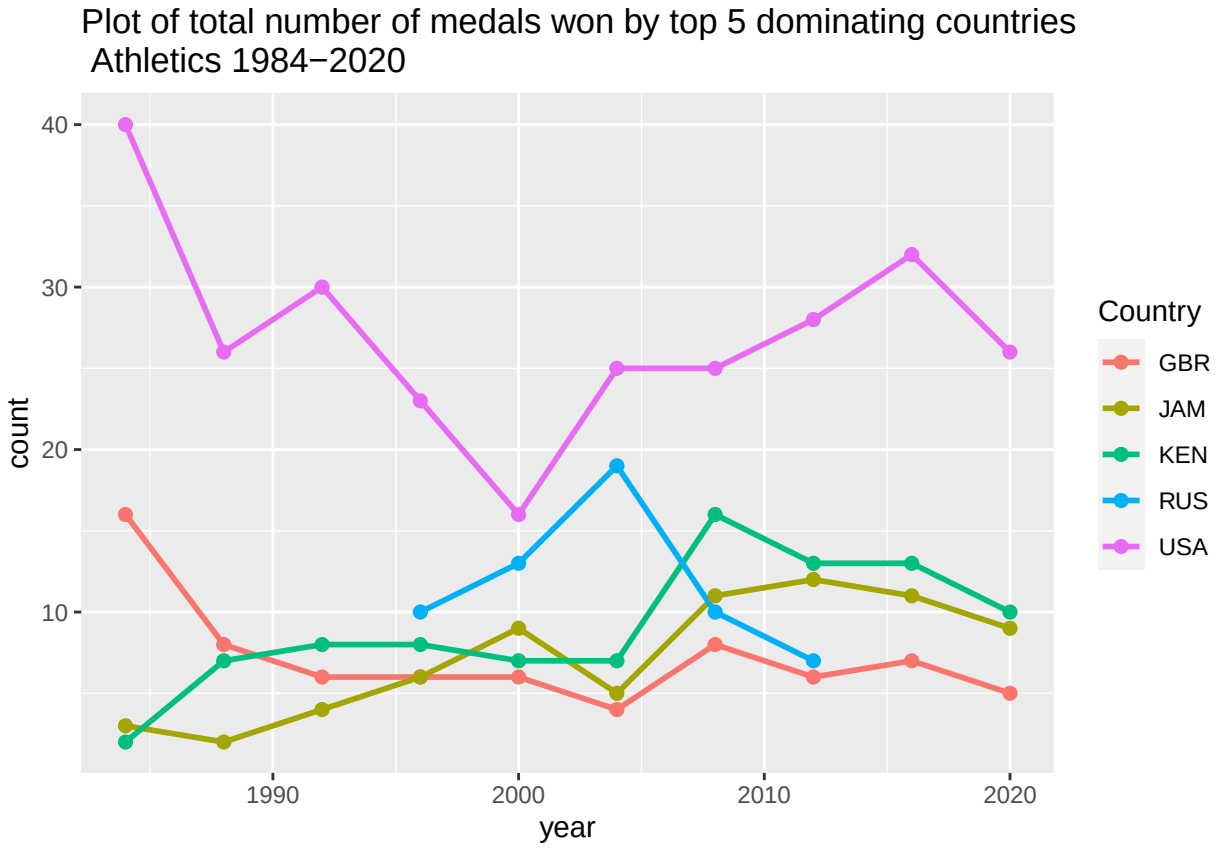
**Plots of medal distribution of countries over years**

**Which countries dominates in which sports?** Finally, we want to know the most dominating countries in each genre of the games and how their medal count varies over years. To do this, first for each sports, we have determined top five dominating countries over the years and yearwise number of medals won by each of those countries are stored in a frame. For example, here is the data obtained for Athletics.

```
## [1] "USA" "KEN" "GBR" "JAM" "RUS"
```

```
## # A tibble: 6 x 3
## # Groups:   year [2]
##    year Country count
##   <dbl> <chr>   <int>
## 1  1984 GBR        16
## 2  1984 JAM         3
## 3  1984 KEN         2
## 4  1984 USA        40
## 5  1988 GBR         8
## 6  1988 JAM         2
```

Here, the top 5 dominating countries in Athletics in last 10 summer Olympics are - USA, Kenya, Great Britain, Jamaica and Russia. Now we plot the number of medals won by these countries over last 10 olympics to see their progress in that event.




## Shiny App

In the Shiny app all the visualizations have been categorised into two different sectors - One is for individual years and another one is for individual events.

- When **plots for individual years** is selected, we can choose any of those 10 years and get 4 different types of plots based on that particular year.
  - Medal distribution of different countries on world map
  - Barplot of medal winners for different kinds of sports
  - Barplot of medal winners based on gender
  - Scatterplot of GDP with the number of medals in that particular year. Here each countries are mentioned by different colours.
- Next **plots for individual events** is selected. In that case we plotted the graphs over the years. We can select a particular event for a particular year. For that particular event we plotted three graphs.
  - Age, Height and Weight distributions are drawn for each event.Scatter plot between every pair of the factors are also drawn.
  - 3 boxplots of Age ,Height and Weight of the medal winners for different events.
  - Scatterplot of number of medals won by the top 5 dominating countries for each sport over the years 1984 - 2020.

## Conclusion

It is to be noted that the whole project is based on Exploratory Data Analysis, i.e., all the inferences we got are completely based on the insights obtained from different kinds of plots and no statistical models have been used to draw conclusions.

However, we can atleast confirm this much that if a player is from a host country or from a country with high GDP and also in optimal range of age and body measurements (height and weight) suitable for a particular sport, he/she has a very high chance of winning a medal in that event.

However, this relation is not that much linear as it seems. Lots of factors are there which haven't been taken into consideration. Notable players like Michael Phelps or Usain Bolt have won lots of medals in different Olympics, irrespective of most of the factors we have considered. End of the day, individual talent, hardwork and practice are the keys of success, whatever the game is.

## Acknowledgement

We would like to express our sincere gratitude to Dr. Dootika Vats for providing an opportunity to work on this project. Her guidance and invaluable advice helped a lot during the course of the project. Also, thanks to our team for immense involvement throughout the work-period.

## References

1. http://www.olympedia.org/
2. https://www.macrotrends.net/countries/ranking/gdp-gross-domestic-product
3. https://www.wikipedia.org/