Olympics Data Analysis

Ranjit Prasad, Chhavi Viswakarma, Arnav Singla, Roy Shivam

1. Introduction

1.1 About

The most famous of sporting competitions are the Olympic Games, which originally took place in ancient Greece between 776 BC and 393 AD. The first modern Games were held in 1859. The Games have come to represent the ideal of sporting excellence -- the "Olympian spirit" of pure athletic competition, untrammelled by worldly considerations.

Olympics has a rich history, spanning from 1896 till 2018, and has been a part of history. So, it is an interesting topic to see how the historical events have affected the specifics of Olympics and how it has been changing till date. Hence, this report attempts to build around the following questions, with some connection to historical happenings.

- *The first Olympic Games had achieved major importance in Greece by the end of the 6th century BCE.*
- *They began to lose popularity when Greece was conquered by Rome in the 2nd century BCE, and the Games were officially abolished about 400 CE because of their pagan associations.*
- *The Olympics were revived in the late 19th century, with the first modern Games being held in Greece is 1896.*
- *The Summer Olympic Games and Winter Olympic Games are each held every four years.*
- *After 1992, when both a Summer and Winter Games were held, they have been held on a staggered two-year schedule so that the Olympic Games occur every two years in either summer or winter.*
- *The International Olympic Committee chooses the location of each Olympic Games.*
- *The choice is based on applications made by the chief authority of a city, with support of the national government.*
- *In individual Olympic events the award for first place is a gold medal, for second place a silver medal, and for third place a bronze medal.*
- *Diplomas are awarded for fourth through eighth places, and all competitors and officials receive a commemorative medal.*
- *Winning a bid to host the Olympic Games has been considered a major boon for any city, but not all agree.*
- *Proponents believe that hosting the Olympic Games can increase valuable tourism, boost local economies, and grow a host country's global trade and stature.*
- *Others maintain, however, that the Olympics are a financial drain on host cities and force them to create expensive infrastructure and buildings that fall into disuse.*

1.2 Motivation behind the project

- *Olympics is the most prestigious sports event and it's a pride for the athletes to participate in this event and win medals for their nation.*
- *Loads of data was available online on Olympics, so we could do a lot of things and play around to figure out the various factors that could be responsible for winning more medals and make some interesting visualizations.*

1.3 Interesting Questions

Here we would like to pose some captivating and enthralling questions which we can answer using our data-sets and the analysis we performed on it.

- **Q** *Is there any sort of bias or correlation between being a host country and increasing a country's medal count?*
- **Q** *Is the total medal count for each nation linked to it's literacy rate, per capita alcohol consumption, prices of beer, year of independence or any other sort of factor which we might be able to think about?*
- **Q** *What are some major factors contributing to a country's success at the Olympics??*
- **Q** *How does the Medal count of country vary over the years??*
- **Q** *What is the Age, Weight and Height distribution of various athletes taking part in Olympics from different nations??*
- **Q** *What has happened to the participation of Nations over the years and has the number of events over the years increased ? If YES then what is the rate.*
- **Q** *Has Olympics become a popular among people??

2. Data

2.1 Dataframes

To get started, we collected the total count of medals won by different countries over the years(1896 - 2020). We also collected the various other factors of the countries like literacy rate, population, life expectancy, BMI, per capita alcohol consumption, etc. These demographic details were the latest we could find, i.e., for year 2020. The data containing the list of athletes and their sex, age, height, weight, medals won was collected.

Overall, we primarily had three datasets:

i) ***FinalYearwiseMedals.Rdata*** : This data frame contained the year-wise(gold, silver, bronze and total) medals won by various countries over the years. Alse the host nation for that year is also mentioned in a column. This data contains 2071 observations with 8 variables. The features of this dataframe are as follows:

- ***year*** *: The years in which the Olympic games were organised.*
- ***Rank*** *: Rank of the nation in the specified year.*
- ***Nation*** *: Name of the country.*
- ***Gold*** *: Number of gold medals won by the specified country in a particular year.*
- ***Silver*** *: Number of silver medals won by the specified country in a particular year.*
- ***Bronze*** *: Number of bronze medals won by the specified country in a particular year.*
- ***Total*** *: Total number of medals medals won by the specified country in a particular year.*
- ***Host*** *: Name of the country that hosted the event in the specified year.*

Few observations of this dataframe is as follows:

```
##   year Rank        Nation Gold Silver Bronze Total HostNation
## 1 1896    1 United States   11      7      2    20     Greece
## 2 1896    2        Greece   10     18     19    47     Greece
## 3 1896    3       Germany    6      5      2    13     Greece
## 4 1896    4        France    5      4      2    11     Greece
## 5 1896    5 Great Britain    2      3      2     7     Greece
## 6 1896    6       Hungary    2      1      3     6     Greece
```

ii) ***FinalTotalMedals.Rdata*** : This data frame contained the total number of medals (gold, silver and total) won by different nations over the history of olympics. This dataset also contains various demographic details of the nations like population(scaled down by 1000), literacy rate, life expectancy, BMI, alcohol consumption, etc. We could find the yearwise data of these demographic details of various countries, so we got the most recent (as of 2020) data we could find. Also, the year of independence of the various nations were added in this dataset. There were a total of 2071 observations and 8 variables in this dataset. The features of this dataframe are as follows:

- ***Nation*** *: Name of the country.*
- ***Gold*** *: Cumulative number of gold medals won by the country over the years.*
- ***Silver*** *: Cumulative number of silver medals won by the country over the years.*
- ***Bronze*** *: Cumulative number of bronze medals won by the country over the years.*

- ***Total*** *: Total number of medals won by the country over the years.*
- ***Literacy*** *: Literacy rate of the country.*
- ***pop*** *: Population of the country(scaled down by 1000).*
- ***lifeExp*** *: Lifr Expectancy.*
- ***BMI*** *: Average BMI of the nation.*
- ***alchol_consum*** *: Rate of alcohol consumption.*
- ***forest_area*** *: Proportion of forest cover in the country.*
- ***BeerPrice*** *: Beer price in the country.*
- ***Indep_year*** *: Year of independence.*
- ***suicide_rate*** *: Suicide Rate*
- ***unemployment_rate****(%) : Unemployment rate.*

Few observations of this dataframe is as follows:

```
##         Nation Gold Silver Bronze Total Literacy       pop lifeExp
## 1      Algeria    5      4      8    17  79.6084 44903.225   77.06
## 2    Argentina   21     26     30    77  98.0900 45510.318   76.81
## 7   Azerbaijan    7     14     28    49  99.8053 10358.074   73.12
## 8   Azerbaijan    7     14     28    49  99.8053 10358.074   73.12
## 10    Barbados    0      0      1     1  99.7000   281.635   79.31
## 11     Belarus   21     37     47   105  99.7220  9534.954   74.23

##     BMI alchol_consum forest_area BeerPrice Indep_year suicide_rate
## 1  26.2           0.9        0.82      1.53       1962          2.5
## 2  27.7           9.8       10.44      1.16       1816          8.4
## 7  27.4           0.8       13.15      2.03       1918          4.1
## 8  27.4           0.8       13.15      2.03       1991          4.1
## 10 28.7           9.6       13.92      1.17       1966          0.6
## 11 26.6          11.2       43.21      0.71       1991         21.2
```

iii) ***height_age_weight.Rdata*** : The sex, age, height weight and Nation of the various athletes is contained in this dataset. This is the largest dataset we collected that contained information about 206165 athletes with 5 variables. The features of this dataframe are as follows:

- ***Sex*** *: Sex of specific the player.*
- ***Age*** *: Age of the athlete.*
- ***Height*** *: Height of the athlete.*
- ***Weight*** *: Weight of the athlete.*
- ***Nation****: Name of the country, which the athlete represents.*

```
##   Sex Age Height Weight      Nation
## 1   M  24    180     80       China
## 2   M  23    170     60       China
## 5   F  21    185     82 Netherlands
## 6   F  21    185     82 Netherlands
## 7   F  25    185     82 Netherlands
## 8   F  25    185     82 Netherlands
```

2.2 Data extraction

We used the "rvest" library of R to extract the data from various websites.

- *Firstly, we got the all-time total medals won by the countries from the following website: https://en.wikipedia.org/wiki/All-time_Olympic_Games_medal_table*

- *The point table for the year 2020 was extracted from the following website: https://olympics.com/en/olympic-games/tokyo-2020/medals Similarly the points table for different years were collected and merge in a single dataframe.*

- *The following websites were scrapped to extract the various demographic details of the countries:*

  - *https://worldpopulationreview.com/country-rankings/literacy-rate-by-country*
  - *https://en.wikipedia.org/wiki/List_of_countries_by_body_mass_index*

The data which contained the names, sex, age, height and weight of various athletes participated in olympics over the years was downloaded from the following website in .csv format and cleaned further.

- *https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results?select=athlete_events.csv*

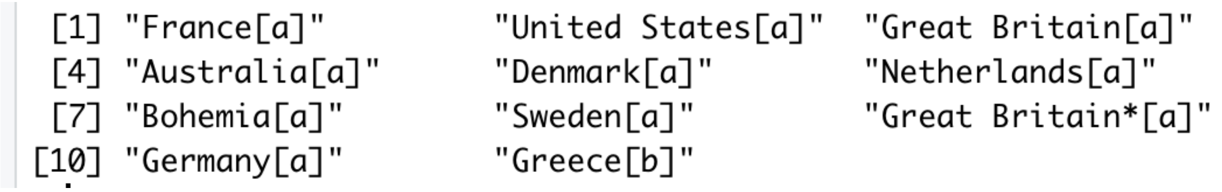
2.3 Data Cleaning

Now lets move to a very crucial part of data analysis, that is data cleaning.We have loads of Data around us, but it is not useful to us if it is not in an organized and regular manner on which we can operate and draw conclusion.

- Since we used multiple factors for each nation, each dataframe was merged together along the Nation using the "merge" function.

- Our data sets had **2 major issues** with respect to data cleaning:

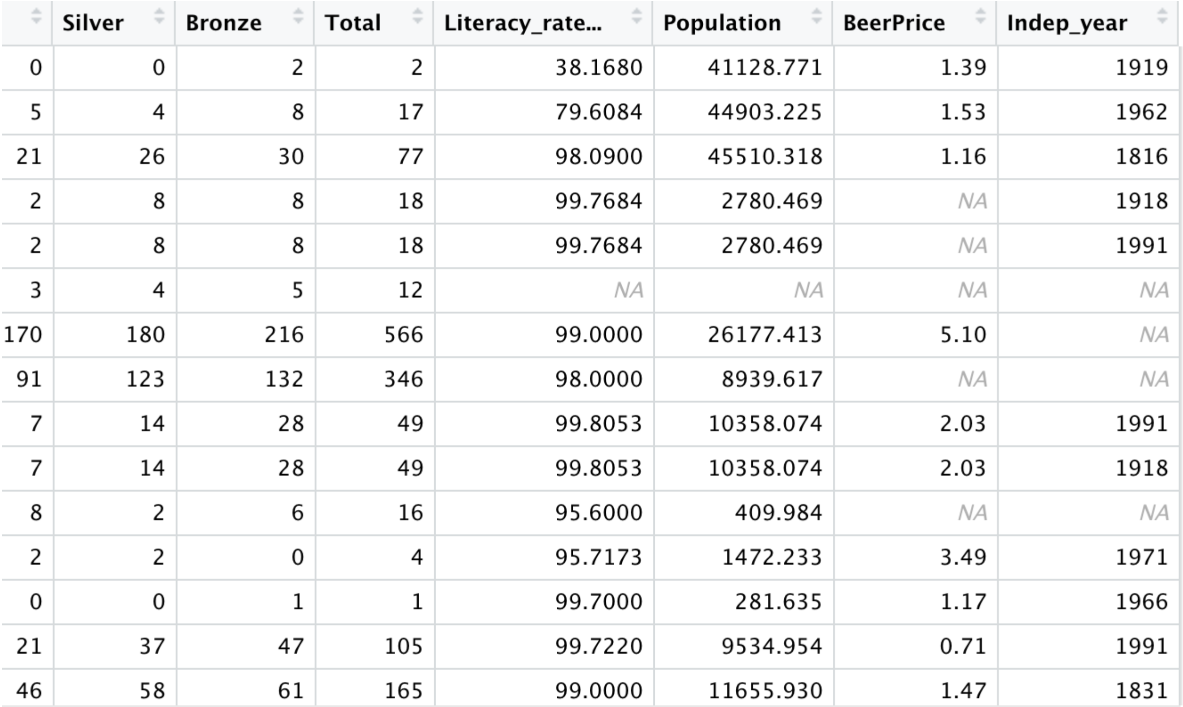
  1. **Irregularities in names of countries :**



```
 [1] "France[a]"         "United States[a]"  "Great Britain[a]"
 [4] "Australia[a]"      "Denmark[a]"        "Netherlands[a]"
 [7] "Bohemia[a]"        "Sweden[a]"         "Great Britain*[a]"
[10] "Germany[a]"        "Greece[b]"
```

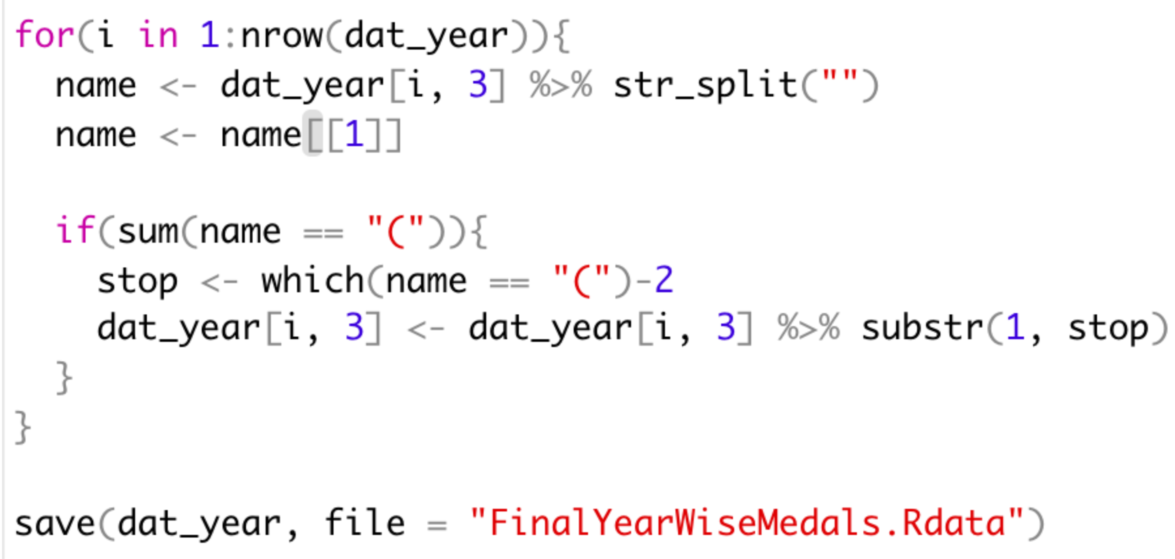
Figure 1: had to clean such names

  2. **Na values corresponding to some factors :**



	Silver	Bronze	Total	Literacy_rate...	Population	BeerPrice	Indep_year
0	0	2	2	38.1680	41128.771	1.39	1919
5	4	8	17	79.6084	44903.225	1.53	1962
21	26	30	77	98.0900	45510.318	1.16	1816
2	8	8	18	99.7684	2780.469	NA	1918
2	8	8	18	99.7684	2780.469	NA	1991
3	4	5	12	NA	NA	NA	NA
170	180	216	566	99.0000	26177.413	5.10	NA
91	123	132	346	98.0000	8939.617	NA	NA
7	14	28	49	99.8053	10358.074	2.03	1991
7	14	28	49	99.8053	10358.074	2.03	1918
8	2	6	16	95.6000	409.984	NA	NA
2	2	0	4	95.7173	1472.233	3.49	1971
0	0	1	1	99.7000	281.635	1.17	1966
21	37	47	105	99.7220	9534.954	0.71	1991
46	58	61	165	99.0000	11655.930	1.47	1831

- Let's see how we dealt with these issues:
  - To clean the name of countries.The **stringr** package was used. first we split the name into characters and then depending on the present irregularity like "*"/"()" / "[]" we make a subset of the data-set then clean it accordingly.



```
for(i in 1:nrow(dat_year)){
  name <- dat_year[i, 3] %>% str_split("")
  name <- name[[1]]

  if(sum(name == "(")){
    stop <- which(name == "(")-2
    dat_year[i, 3] <- dat_year[i, 3] %>% substr(1, stop)
  }
}

save(dat_year, file = "FinalYearWiseMedals.Rdata")
```

  - while merging data on different factors based on the nation some null entities were introduced in the dataframe, so of them had to be filled manually and rest were dropped using "na.omit" ### 2.4 Key Questions of Interest

The purpose of our analysis was to answer some, if not all, of the following questions:

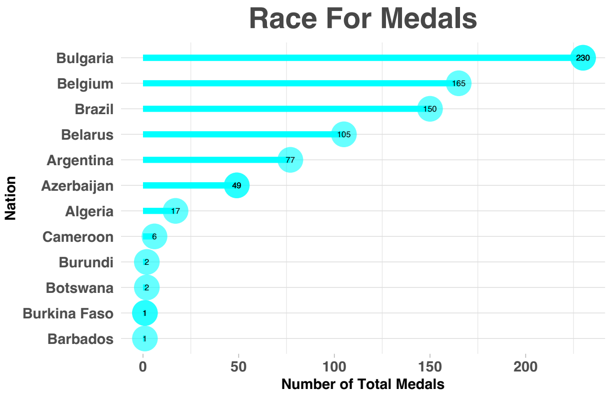
- *Which are the top performing countries in Olympics and does the various demographic factors like life expectancy, literacy rate, year of independence, etc. affect the medal count of a country?*

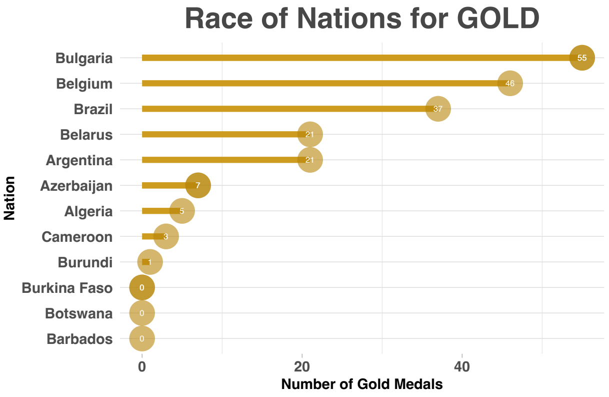
- *Has popularity of Olympics increased over the years ?*
- *What is the age, height and weight distribution of the athletes ?*
- *Does hosting the event increase the chances of winning more medals ?*

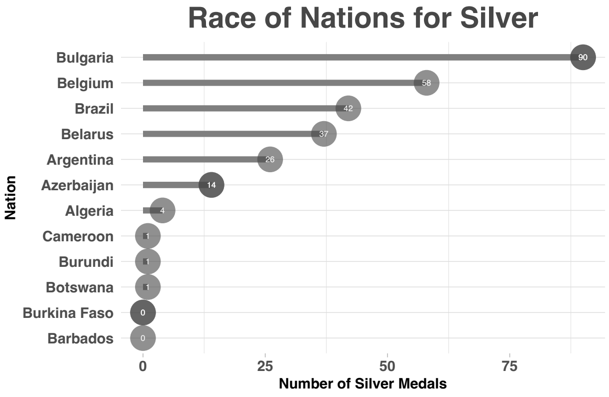
3. Analysis and visualization

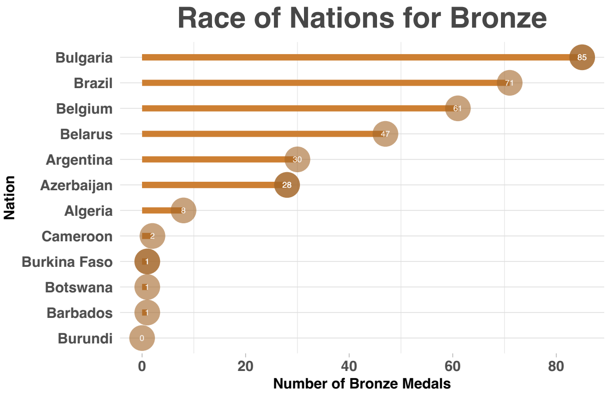
3.1 Race for Medals

- *First of all We we wanted to analyze where the various nations lie in terms of the number of medals like gold/silver/bronze and the total Number of Medals which each nation has won till date.*
- *So basically the following graph depicts that. You can select the Nations you want to see on the leader-board and then we'll depict the ranks for each type of medals.*



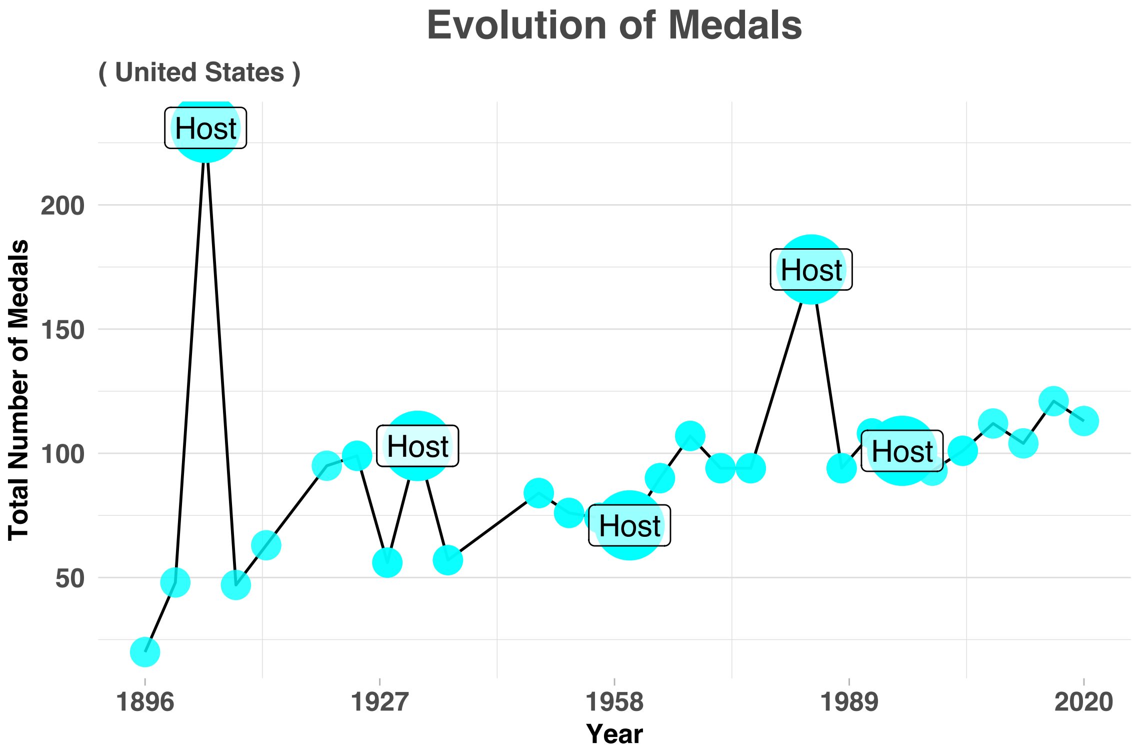






3.2 Evolution of Medals

- *Next we wanted to study the year-wise distribution of medals for each Nation and the important thing we wanted to see was does* **hosting the Olympics** *for a particular nation implies that it will win more medals.*
- *In the following graph we have used* **United States** *as an example to depict the above Hypothesis. We can see that among st all the Peaks most of the peaks are those in which US has hosted the Olympics. So it might be the case that the Olympians feel more confident in their home-ground or there is some other sort of bias.*
- *This was observed for a few other countries as well like Greece, France, etc.*



3.3 Various Factors Used

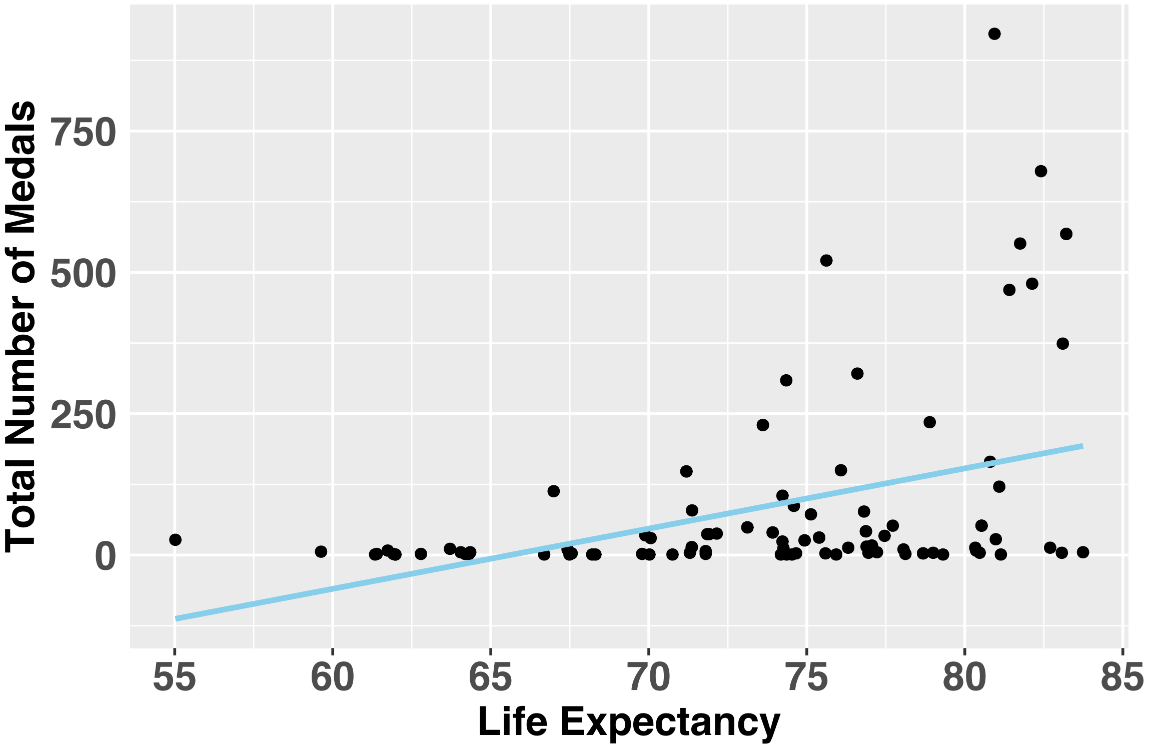
- *Next we are looking at the plots of Total Medals of each country against the factor which was used to calculate the correlation.*
- *Here the line Drawn in each plot is representing the line of best fit according to a linear model.*
- *The 4 Factors which had highest correlation were as follows:*

Note: Here the "total number of medals" means the cumulative sum of gold, silver and bronze medals which a country has won over time.

i)**Life Expectancy** :

This factor was kind of obvious since higher life expectancy implies higher fitness levels and more fit people will definitely win more medals in the Olympics.Thus, Life expectancy showed **positive correlation.**

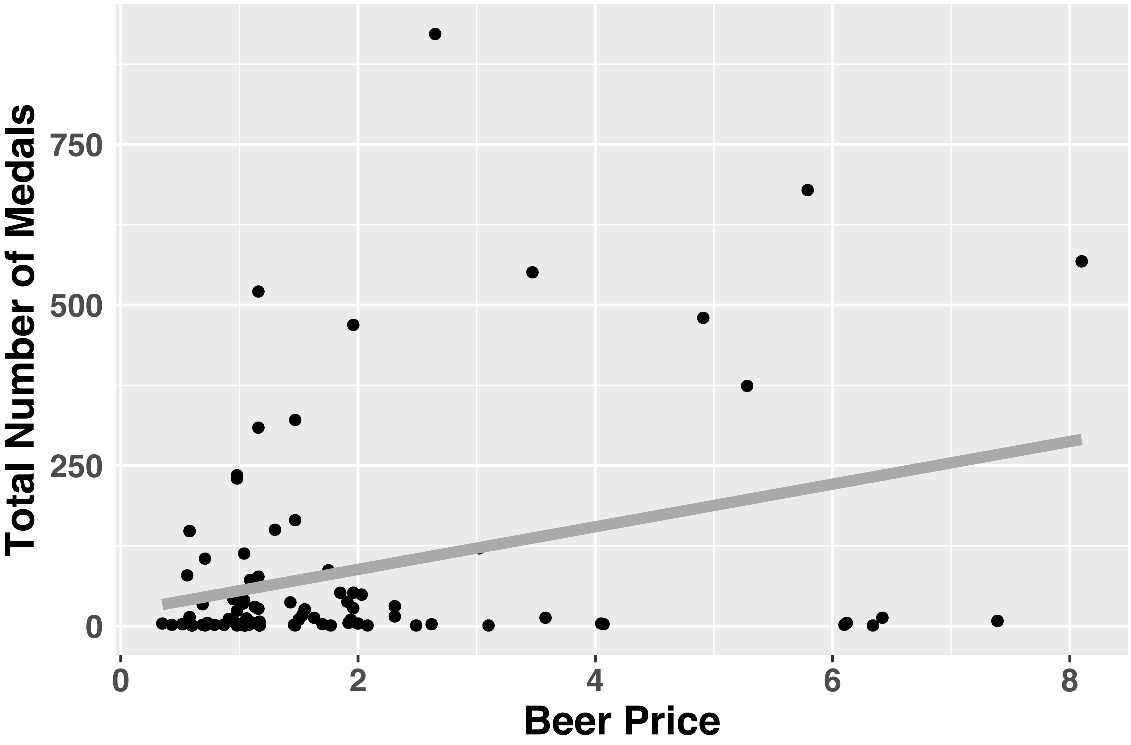
```
## 'geom_smooth()' using formula 'y ~ x'
```



ii)**Price of Beer (in USD) per 500ml** :

- *This factor was pretty amusing for us as well. The important thing to note is here is that we got correlation but we also know that correlation does not imply causation.*
- *So if we consider that, then there might be a confounding variable involved which is affecting both the factors under consideration, .ie Prices of beer and the total number of medals.*
- *On the other hand if consider causation we got to the conclusion that if the beer is cheap, people will drink more beer and will be happy and happy people tend to be more efficient in whatever they do, so they will win more medals.Prices of Beer showed* **negative correlation** *with the total number of medals won.*

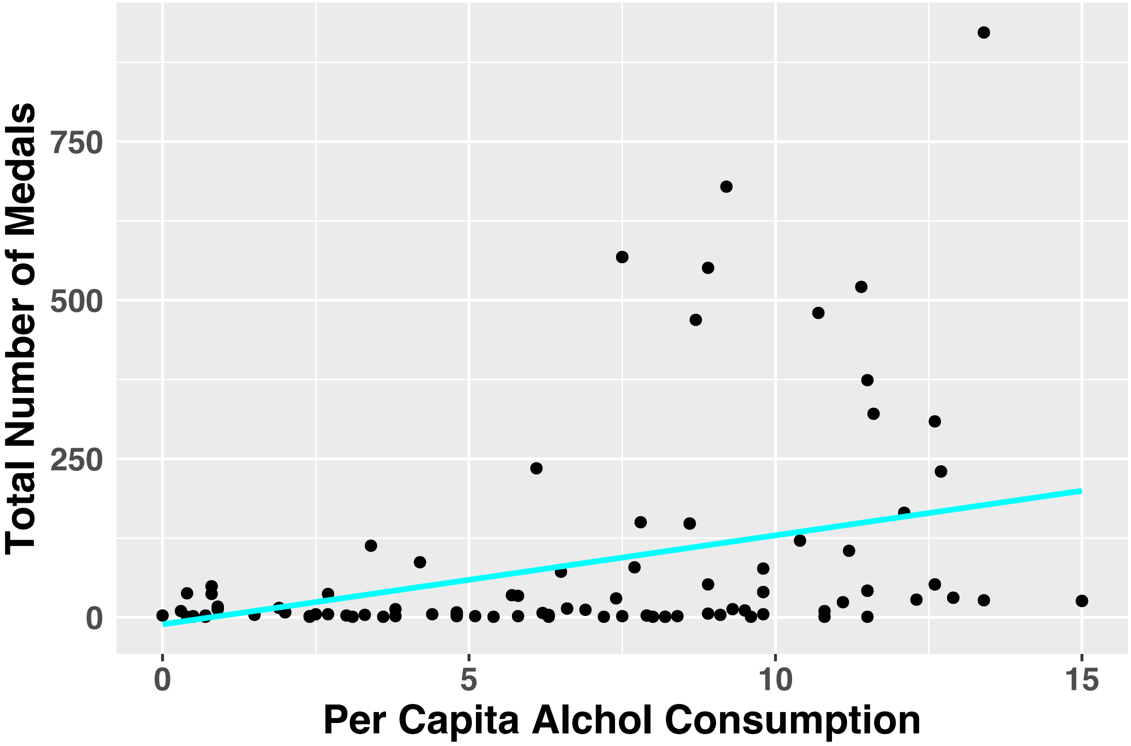
```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



iii)**Per Capita alcohol consumption** :

- *This factor also follows the same discussion as the factor "Price of Beer". It showed* **positive correlation** *with the total number of medals won.*

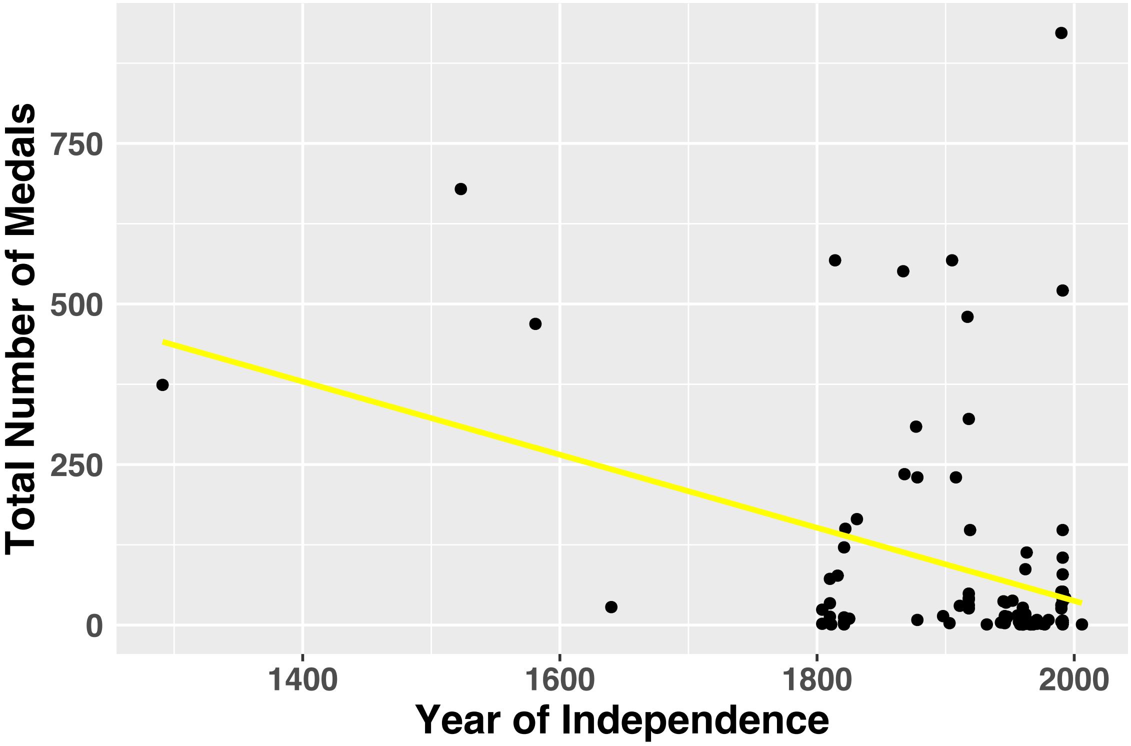
```
## `geom_smooth()` using formula 'y ~ x'
```



iv)**Year of Independence** :

- *We wanted to check whether year of independence had any effect on the leader-board and it actually did.*
- *The countries which got independence later tend to be on the bottom of the leader-board and vice-versa. Thus, Year of Independence showed* **negative correlation** *with the total number of medals won.*

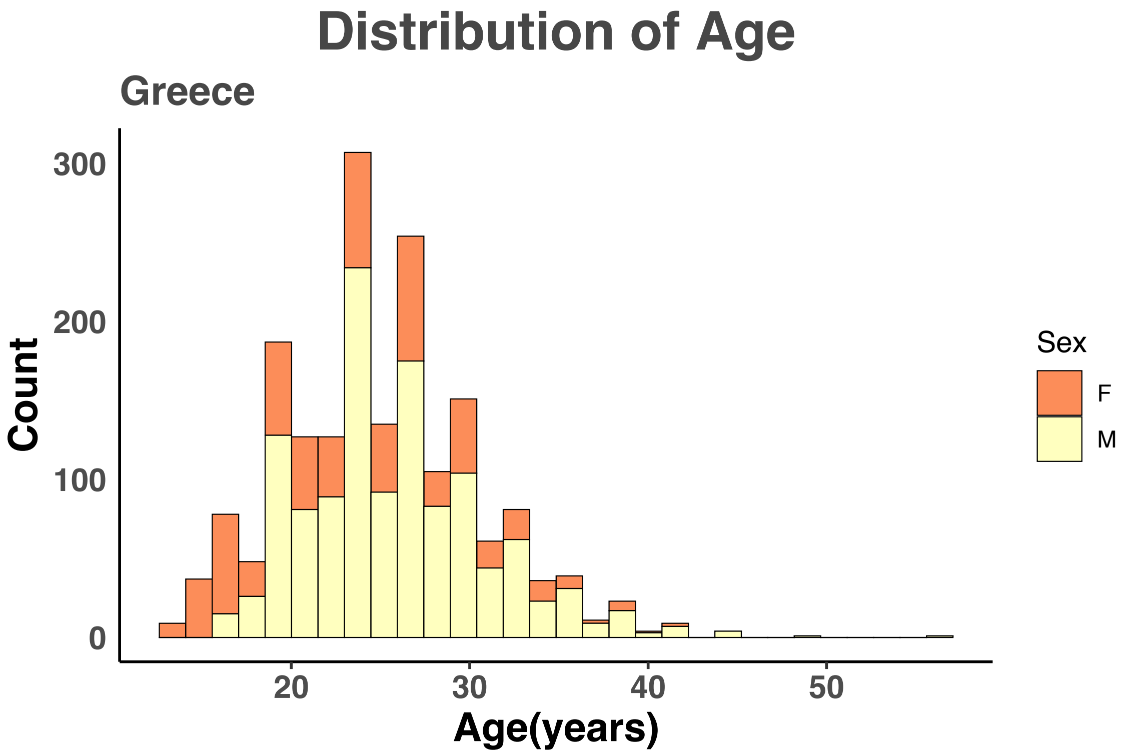
```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



3.4 Distribution of Age, Weight and Height

- *We wanted to study how the various factors like age, weight and height are distributed considering all the athletes in the various events that take place in Olympics.*
- *First let's have a look at the* **Age** *factor:*
- *We found out through our analysis that the average age of Olympians was in the range of* **24 to 28** *, the lowest being 13 and highest being 72(in the shooting category).*
- *In the following plot you can basically select the country for which you want to see the distribution of age and the following plot will be displayed:*

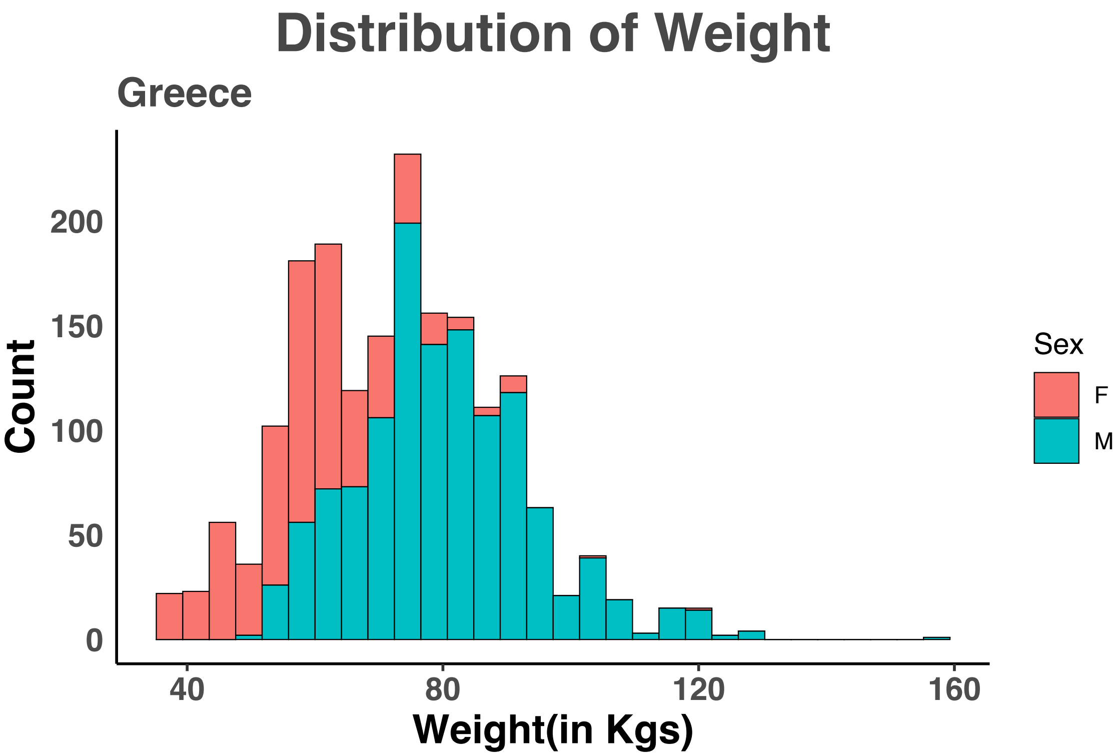
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



Distribution over Weight

- *Now let's look at the* **Weight** *factor:*
- *We concluded through our analysis that the average weight of male Olympian was* **80.1 kg** *and that of female Olympian was* **62.6 kg** *and the overall average came down to* **70.69 kg**.
- *Similar to the above plot you can choose the nation you want to view and we'll give you the distribution of weight for that nation.*

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

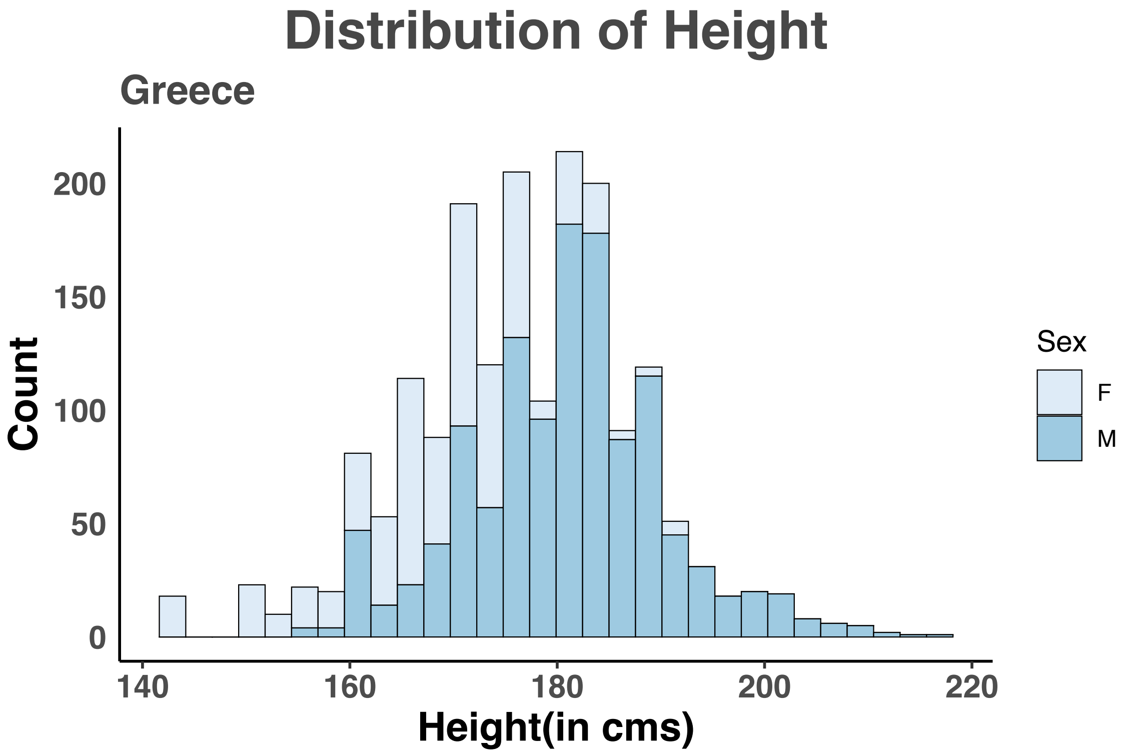


Distribution over Height

- *Now time to analyse the last factor which is***Height***:*
- *What we observed was that the Median Height is* **175 cm***and the 168 cm, 183 cm are the lower and the upper quartile respectively.*
- *The* **maximum** *height for our data-set was observed to be* **226cm** *and the* **minimum** *being* **127cm***. Giving some context the* **average global height** *of a male is* **171cm** *and that of female is* **160cm** *only.*
- *Similar to the above two plots you can choose the nation you want to view and we'll give you the distribution of height for that nation.*

the plot is as follows:

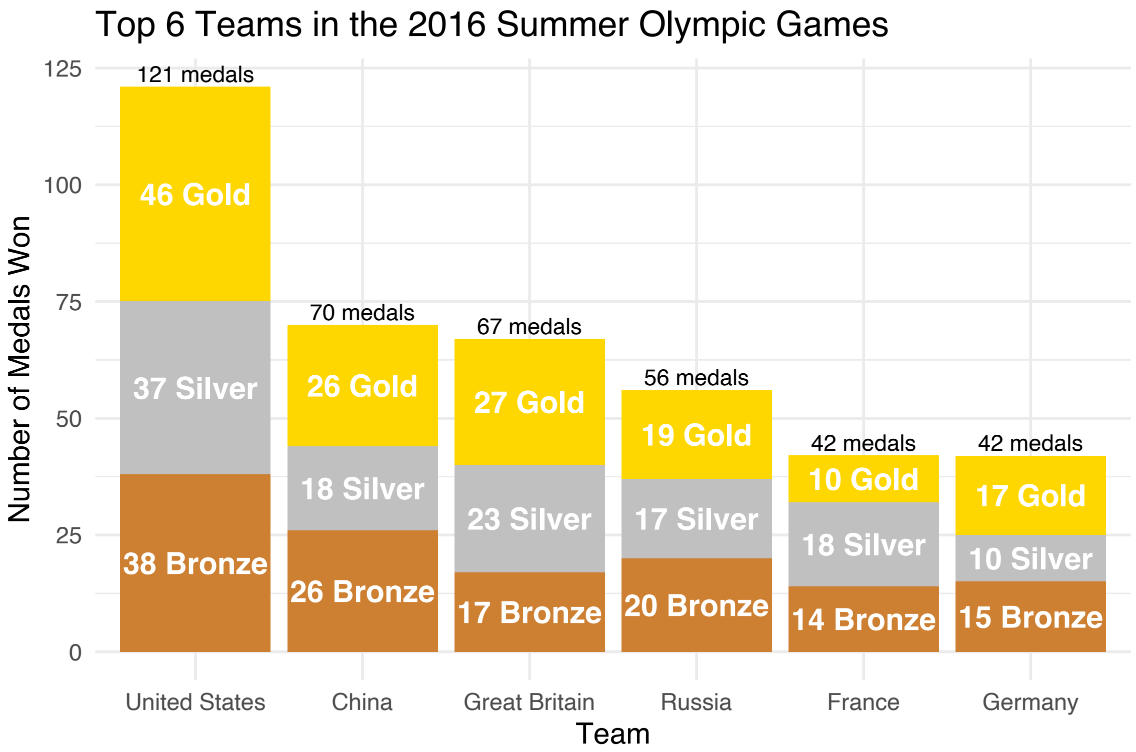
```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



3.5 Some more Graphs

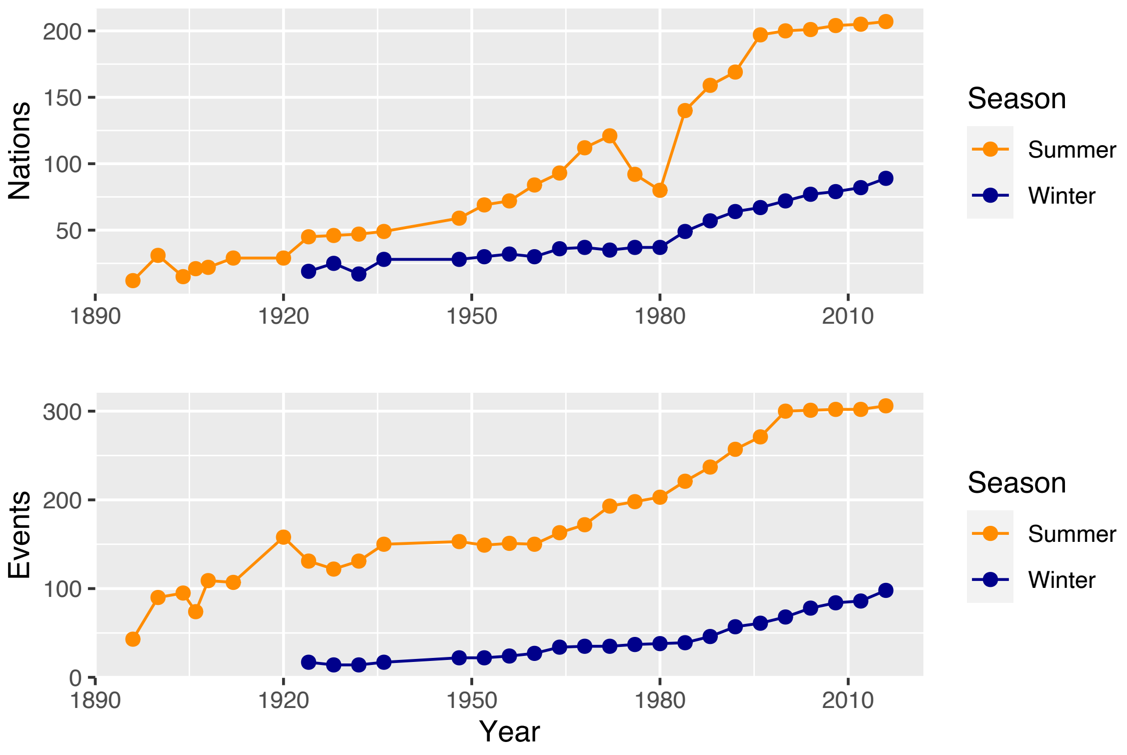
Top performing nations We wanted to get an idea of top performing countries of each olympic session so we plotted a bar graph representing the top 6 countries on the basis of total medals won by the country's athletes.

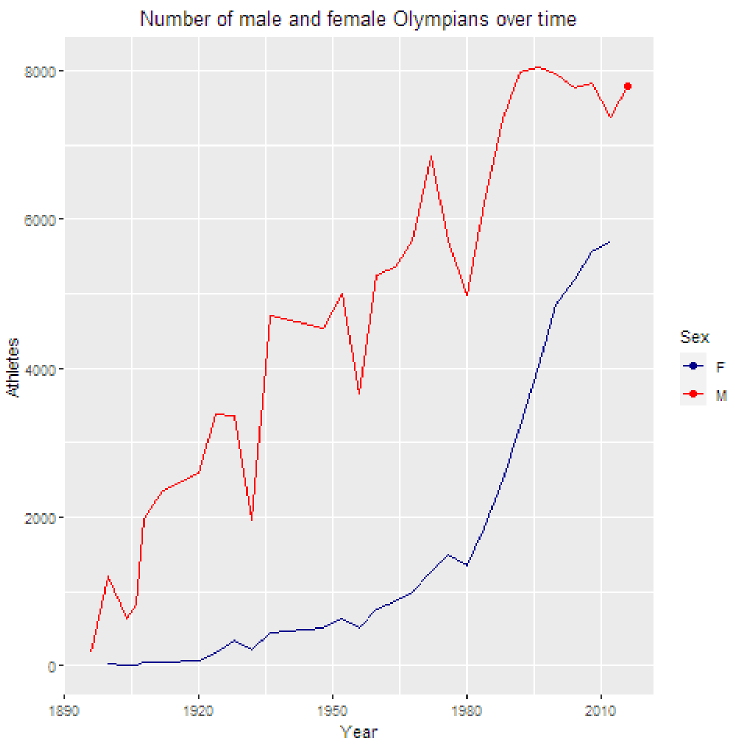
- *The bars of the graph is further sub-divided into number of gold, silver and bronze.*
- *We can see year-wise trend country's and see whether top spots changes or not.*
- *By observing some previous Olympics data we come to conclusion that top countries consistently maintains there performance.*



Olympics Evolution Here we have some line graphs depicting the evolution of the Olympics, like a graph of the total nations participating in the Olympics over the and plot of total events over the year.

- *The graph clearly shows that the number of nations participating in the Olympics has increased over time; even after world war-1, there was an increase in the number of countries participating in the Olympics.*
- *We can also see a peak in the number of events in the Olympics after world war-1 and thought that it was done to get a decent number of participation in sports after the war.*
- *The number of athletes participating over the year shows an overall increasing trend, but there are some dips; for example, the number of participants decreased after world war-2.*





This concludes our **Analysis and Visualization** part of the project.

4. Conclusion

- *It can be clearly seen from the plots that there has been an exponential increase in the participation of countries and athletes. Also the number of events has increased, which shows that the popularity of the Olympics has increased over the years.*
- *It can also be concluded that the countries which got independence early tend to win more medals.*
- *The countries having higher life expectancy, higher per capita alcohol consumption won more medals than others.*
- *And the medal count decreased with increase in unemployment rate.*
- *Germany and Sweden were among the top performing countries.*
- *With few nations we can observe that when they hosted the Olympics their Medals won sought to increase. (USA, Sweden, etc)*
- *The Average age of Olympians was in the range of 24-28.(Highest was 72).*

5. References

- *https://shiny.rstudio.com/gallery*
- *https://worldpopulationreview.com/country-rankings/olympic-medals-by-country*
- *https://triemann.ca/wp-content/uploads/2021/01/Olympic-Analysis_Riemann_Nicol.pdf*
- *https://en.wikipedia.org/wiki/List_of_countries_by_forest_area*
- *https://en.wikipedia.org/wiki/List_of_countries_by_alcohol_consumption_per_capita*
- *https://en.wikipedia.org/wiki/List_of_national_independence_days*
- *https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results*
- *https://en.wikipedia.org/wiki/List_of_countries_by_life_expectancy*
- *https://en.wikipedia.org/wiki/List_of_countries_by_body_mass_index*
- *Beer prices*
- *Suicide Rate*