Assignment 4

Juntong Wei, Qin Xu, Xuanming Liang, Ziang Li

2021-06-04

1 Introduction

This report looks at the data from previous Olympic Games and finds some interesting points. It focuses on the number of medals, the age of the athletes, the gender, the country and the relationship between them. It focuses on the distribution and number of medals, the ranking of countries in terms of the number of medals won, the number of Olympic sports and the relationship between gender, age and medals won. In this report, we use two datasets, named “data” and “dat1”. In “data” has 271,116 obs. and 15 variables. In “dat1” has 184 obs. And five variables. In “dat1”, we mainly use to replace the country code (NOC) with the name of the country, so we use the variables “NOC” and “Country”. In “data” is the main dataset we use, we use “sex”, “age”, “NOC”, “Year”, “sport”, “event” and “medal” these seven variables, and the variables used them in four parts of the study.

2 Analysis

2.1 Medal distribution by country (Xuanming_Liang)

Table 2.1: The medals in different country

Country	NOC	Gold	Silver	Bronze	Total
United States	USA	2638	1641	1358	5637
Germany	GER	745	674	746	2165
United Kingdom	GBR	678	739	651	2068
France	FRA	501	610	666	1777
Italy	ITA	575	531	531	1637
Sweden	SWE	479	522	535	1536
Canada	CAN	463	438	451	1352
Australia	AUS	348	455	517	1320
Russia	RUS	390	367	408	1165
Hungary	HUN	432	332	371	1135

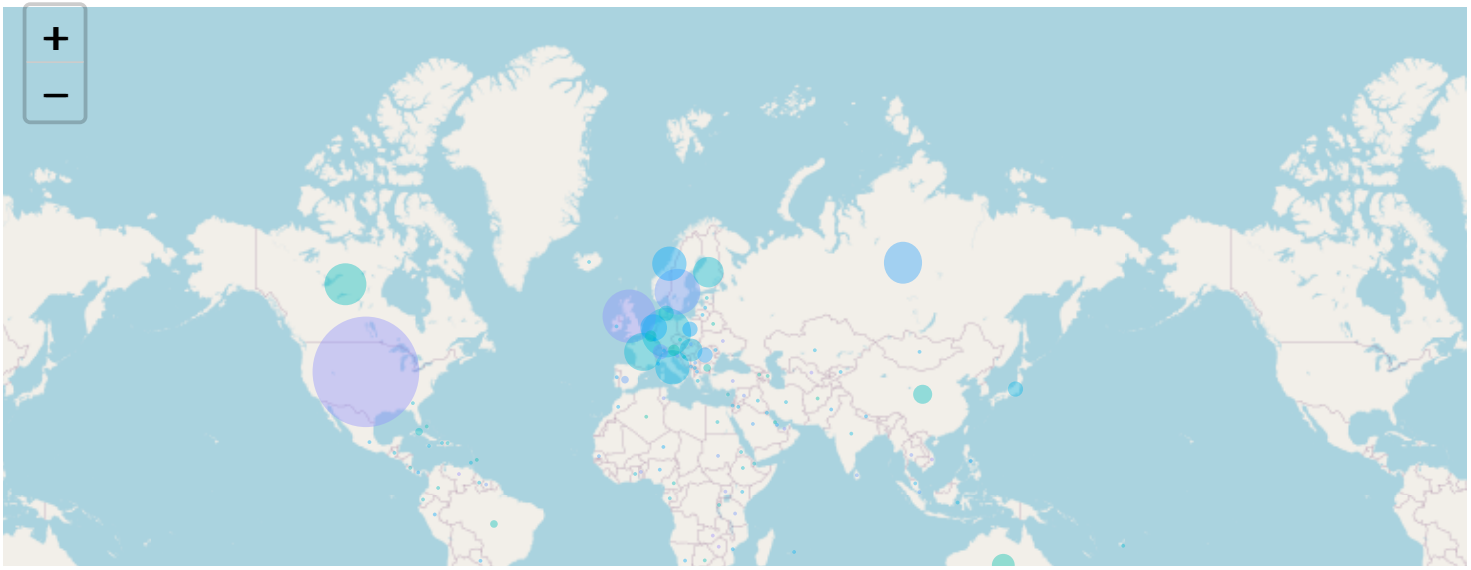

Figure 2.1: Medals distribution in the world wide

The map 2.1 shows the number of medals won by each country in the world at the Olympic Games, including gold, silver, bronze and total. The size of the circles on the map indicates the number of medals won, so it is easy to see that the USA has the most medals. Europe has the highest number of medals, and the density of the circles shows that most European countries have won medals and have accumulated a significant number of medals in total.

2.2 Find out which sport has the largest number of participants and study the distribution of gold MEDALS in different countries over time (Juntong_Wei)

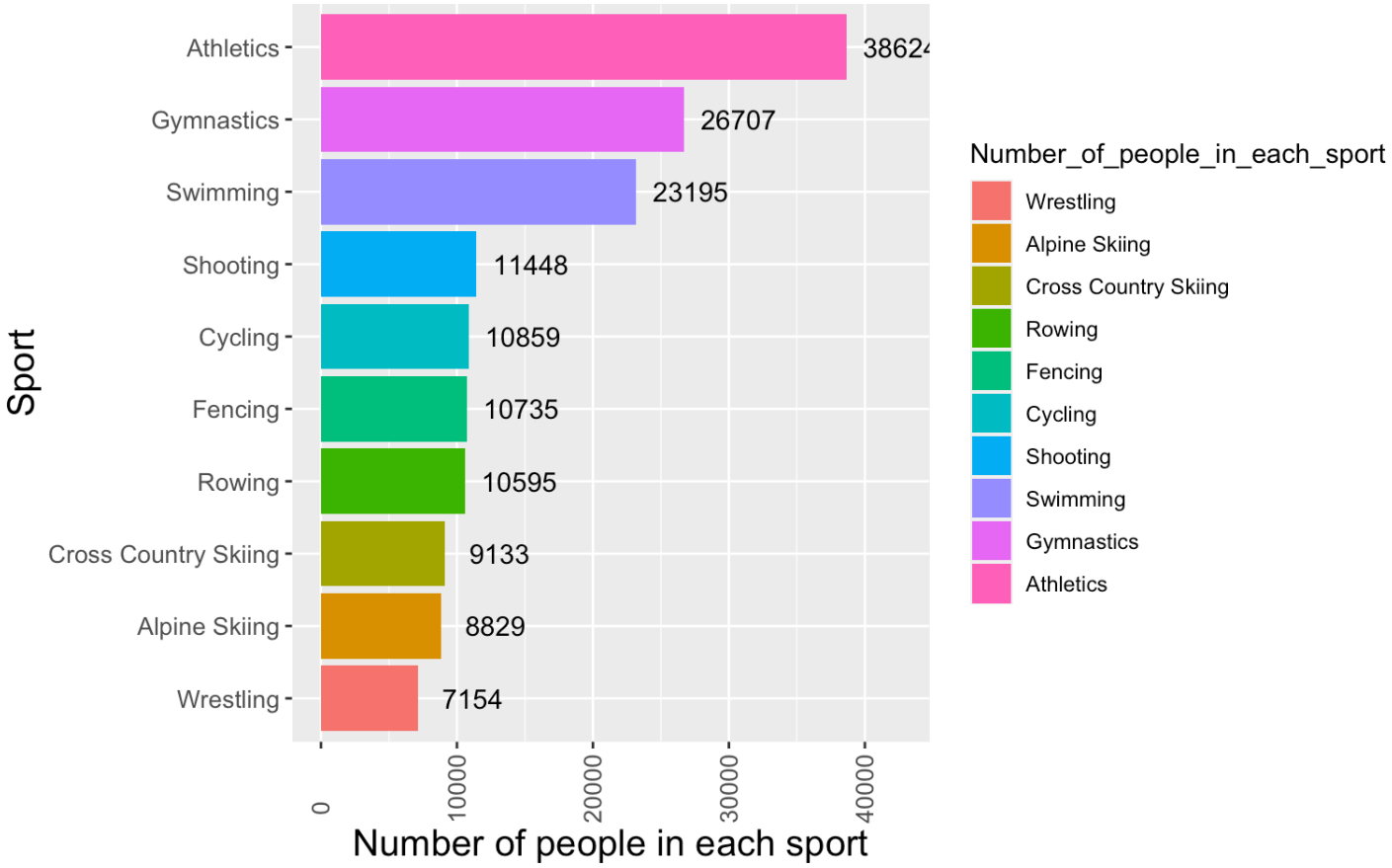

Figure 2.2: Top 10 sports with the most athletes

With the Figure 2.2, It illustrate the number of participants in different sports, and select the top 10 of them, according to the figure, the most popular sports is **Athletics** which is 38,624. the second is **gymnastics** and the third one is **swimming**. the fine thing about this figure is that the number of participate athletes of first 3 sport is far more than any

other sports.

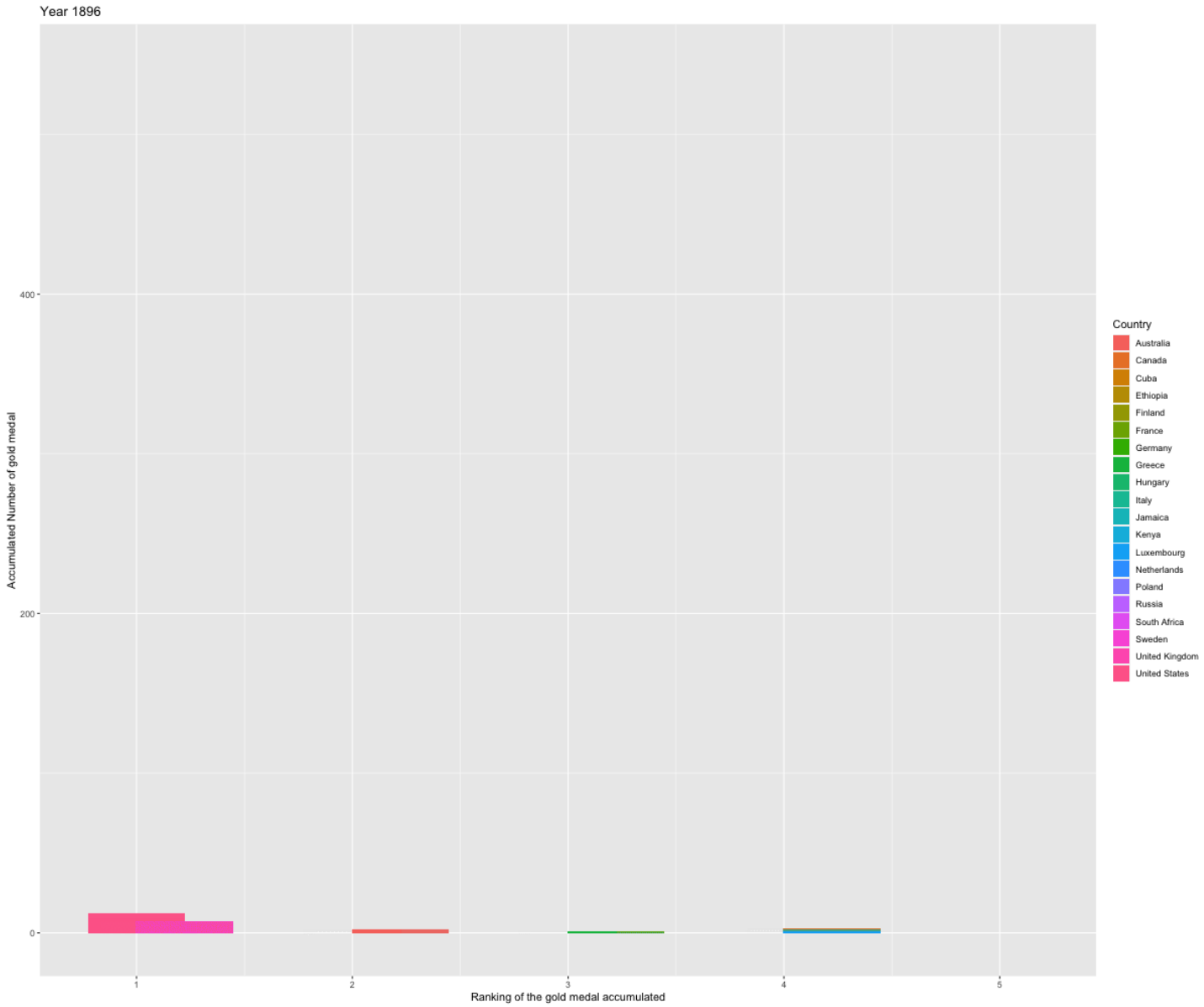

Figure 2.3: In the Athletics sports, the ranking of first 5 contries about Total number of Cumulative gold MEDALS won

Then we focus on the athletics sports, the above Figure 2.3 describes only in the Athletics sports, it shows number of Cumulative gold MEDALS won in each country and ranking them. The x-axis is about first 5 ranking, the y-axis is about the Accumulated Number of gold medal, and the different color means different countries, so there are 2 interesting finding in this GIF,

- The American is always number one in the ranking list except year 1980. That's because the USA did not join the sport meet in that year to boycott the former Soviet Union.
- Sometimes the bar overlapped on the x-axis, which means it shares the same ranking place in this year.

2.3 How about the Medals of top 5 countries allocated in the events (Ziang_Li)

In the section 1, we discussed the number of medals each country has won in the previous Olympic Games. In section2, we analyzed the sports with the largest number of athletes participate in. Hence, in section3, we will compare the five countries with the most medals and the distribution of medals in the five events with the largest number of participants.

From the figure 2.4, it clearly shows that the Medals changing of the Top5 countries in previous Olympic Games. The United States has consistently ranked first in the total number of medals. After 1990, the number of medals in Germany has increased rapidly. And in 1998, Germany ranked second in the total number of medals.

From the figure 2.5, it shows the distribution of the top5 countries medals among the 5 sports with the most participants. The number of medals in the United States is larger than the other four countries and The United States has an absolute advantage in swimming events, especially in the relay race events. The medals of United Kingdom are more distributed in Cycling. In addition, the medals from other countries are very evenly distributed among the five events

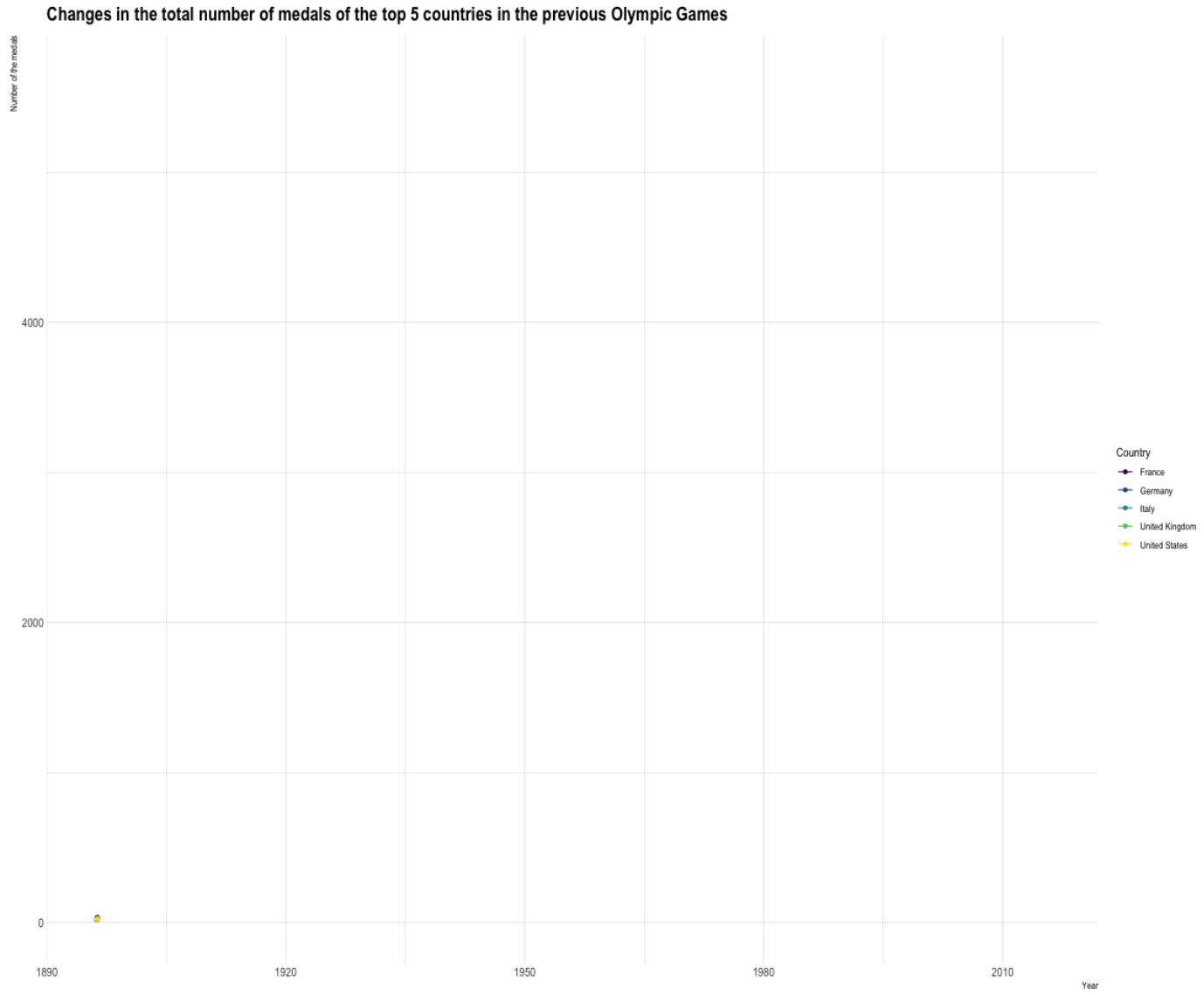

Figure 2.4: Changes in the total number of medals of the top 5 countries in the previous Olympic Games

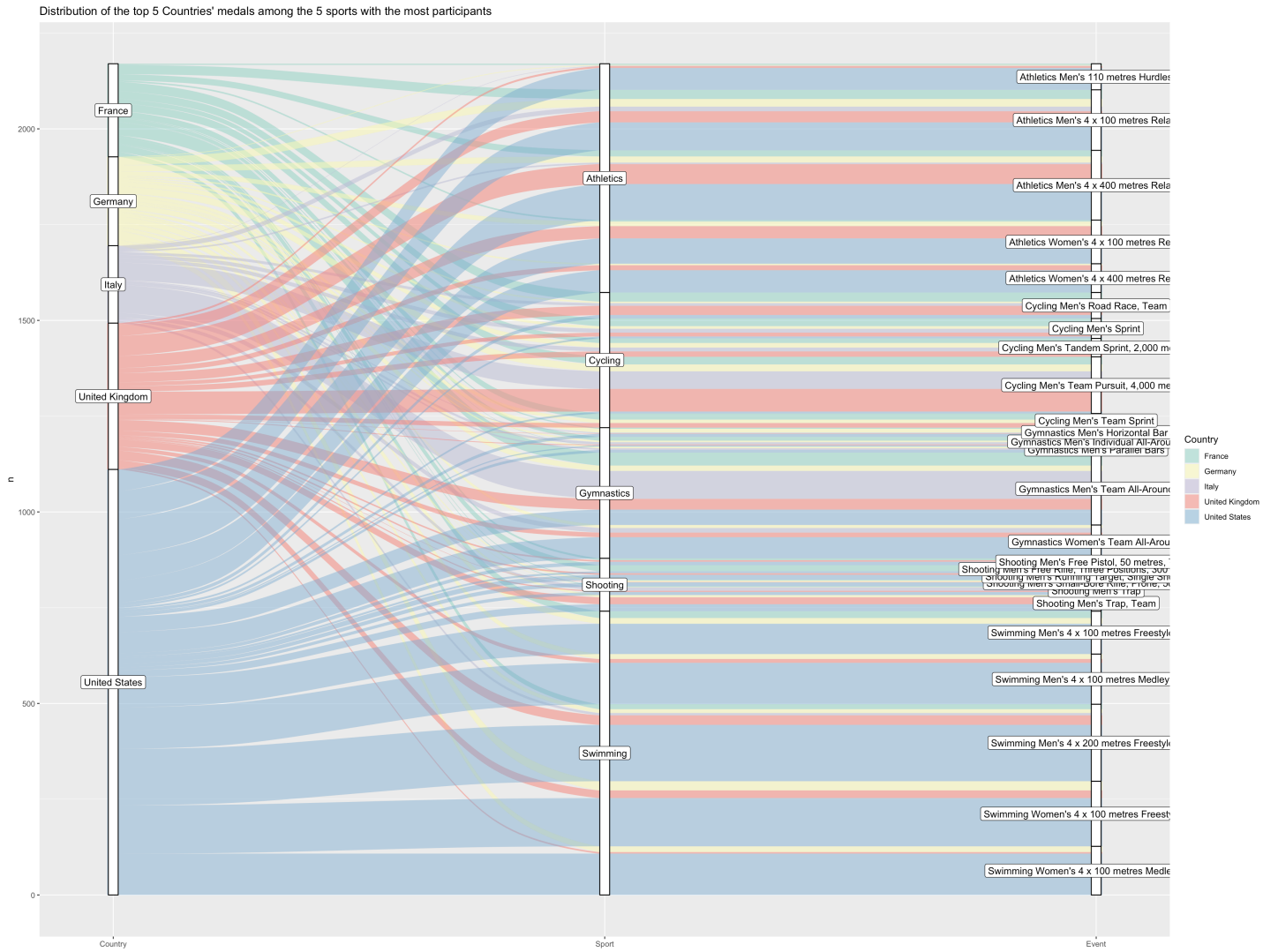

Figure 2.5: Distribution of the top 5 Countries' medals among the 5 sports with the most participants

2.4 Compare the total medal won by different age for both female and male. (Qin_Xu)

According to @ elmenshawy2015rise, age at peak of athletics performance for women have increased in the last 20 years but not for men, hence in this section, we seek to compare the athletics performance of both gender by examining the medal results of athletes who attended Olympics from 1896 to 2016 by age and gender.

```
## # A tibble: 425 x 4
## # Groups:   Sex, Age [138]
##    Sex     Age Medal      n
##    <fct> <dbl> <fct>  <int>
##  1 F        11 Silver     1
##  2 F        11 <NA>      11
##  3 F        12 Bronze     1
##  4 F        12 Silver     3
##  5 F        12 <NA>      28
##  6 F        13 Bronze     2
##  7 F        13 Gold       5
##  8 F        13 Silver     6
##  9 F        13 <NA>     138
## 10 F        14 Bronze    15
## # … with 415 more rows
```

```
## Adding missing grouping variables: `Age`
```

```
## # A tibble: 48 x 4
## # Groups:   Sex, age_group [13]
##    Sex   age_group Medal  number
##    <fct> <fct>     <fct>   <int>
##  1 F     (0,15]    Bronze      4
##  2 F     (0,15]    Gold        3
##  3 F     (0,15]    Silver      5
##  4 F     (0,15]    <NA>        5
##  5 F     (15,30]   Bronze     15
##  6 F     (15,30]   Gold       15
##  7 F     (15,30]   Silver     15
##  8 F     (15,30]   <NA>       15
##  9 F     (30,45]   Bronze     15
## 10 F     (30,45]   Gold       14
## # … with 38 more rows
```

```
## # A tibble: 13 x 3
## # Groups:   Sex [2]
##    Sex   age_group total_medal_in_age_group
##    <fct> <fct>                        <int>
##  1 F     (0,15]                          17
##  2 F     (15,30]                         60
##  3 F     (30,45]                         59
##  4 F     (45,60]                         32
##  5 F     (60,75]                         15
##  6 F     <NA>                             4
##  7 M     (0,15]                          15
##  8 M     (15,30]                         60
##  9 M     (30,45]                         60
## 10 M     (45,60]                         60
## 11 M     (60,75]                         31
## 12 M     (75,90]                          6
## 13 M     <NA>                             6
```

Table 2.2: compare the percentage of thetotal medal won by different age groups for both female and male

Sex	total_medal_by_sex	age_group	total_medal_in_age_group	Percentage
F	187	(0,15]	17	9.090909
F	187	(15,30]	60	32.085561
F	187	(30,45]	59	31.550802
F	187	(45,60]	32	17.112299
F	187	(60,75]	15	8.021390
F	187	NA	4	2.139037
M	238	(0,15]	15	6.302521
M	238	(15,30]	60	25.210084
M	238	(30,45]	60	25.210084
M	238	(45,60]	60	25.210084
M	238	(60,75]	31	13.025210
M	238	(75,90]	6	2.521008
M	238	NA	6	2.521008

In Table 2.2 we compare the total medal won by different age groups for both female and male.

From this table, it shows that female younger athletes age range from 15- 30 has the highest percentage (32.09%) of total medal won compared to other age group, while older female athletes age from 60 to 75 has the least percentage of medal won (8.02%). However, for males athletes, it appears that age group between (15-30),(30-45),(45-60_ share the same percentage(and highest) of total medal won, while older male athletes age form 75 to 90 again has the least percentage of medal won (2.52%).

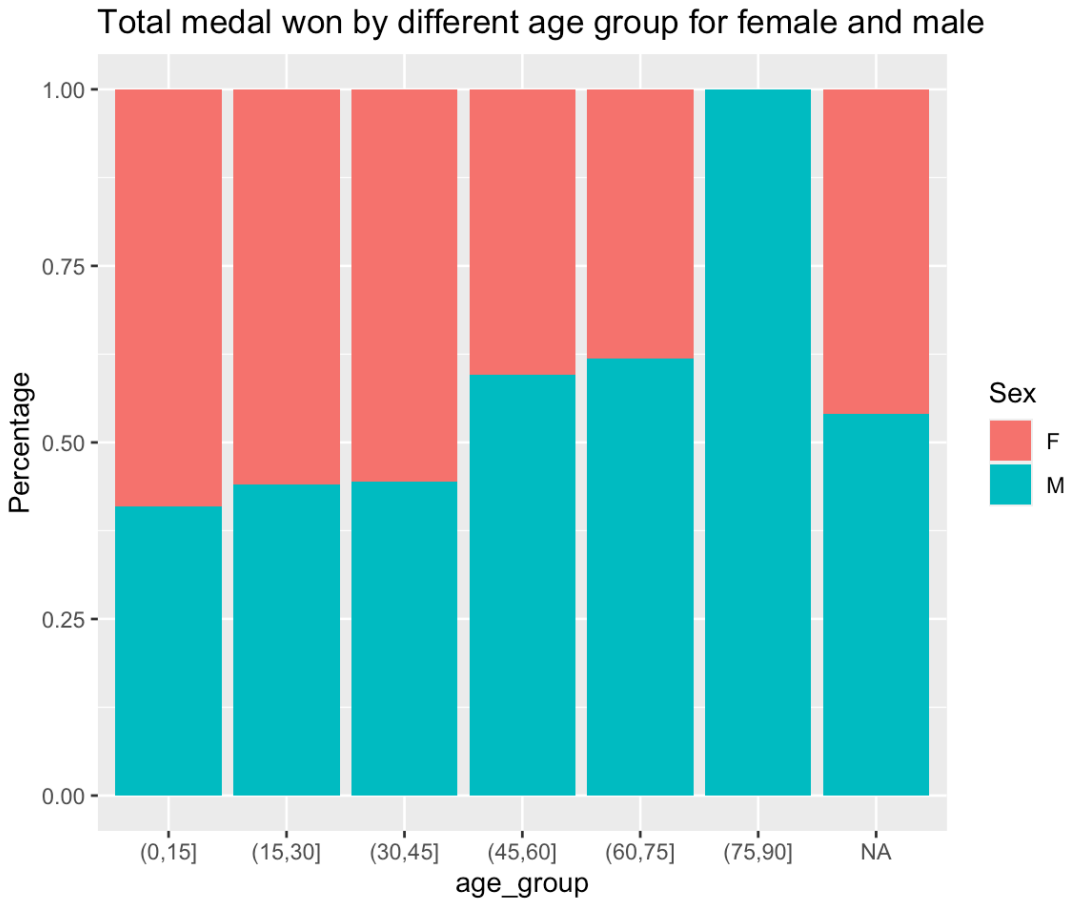

Figure 2.6: The percentage of total medal won in different age group by sex

In Figure 2.6, we have plotted the percentage of the total medal won by different age group and compared in against both gender of male and female.

In Figure 2.6, depicts that for the age groups aged (0-15),(15-30),and (30-45), female athletes is accounted of a higher proportion of the total medal compared to males, However by the age group (45 to 60), males athletes exceed female athletes in the proportion of total medal won and continues to have a higher proportion than female in the older age bracket.

2.5 Comparison of different medal distribution between age group by gender. (Qin_Xu)

```
## # A tibble: 425 x 4
## # Groups:   Sex, Age [138]
##    Sex     Age Medal  number
##    <fct> <dbl> <fct>   <int>
##  1 F        11 Silver      1
##  2 F        11 <NA>       11
##  3 F        12 Bronze      1
##  4 F        12 Silver      3
##  5 F        12 <NA>       28
##  6 F        13 Bronze      2
##  7 F        13 Gold        5
##  8 F        13 Silver      6
##  9 F        13 <NA>      138
## 10 F        14 Bronze     15
## # … with 415 more rows
```

```
## # A tibble: 89 x 6
## # Groups:   Sex, Age [89]
##    Sex     Age Medal number total_silver_medal Percentage
##    <fct> <dbl> <fct>  <int>              <int>      <dbl>
##  1 F        13 Gold       5               3747      0.133
##  2 F        14 Gold      20               3747      0.534
##  3 F        15 Gold      66               3747      1.76
##  4 F        16 Gold     103               3747      2.75
##  5 F        17 Gold     133               3747      3.55
##  6 F        18 Gold     160               3747      4.27
##  7 F        19 Gold     177               3747      4.72
##  8 F        20 Gold     188               3747      5.02
##  9 F        21 Gold     265               3747      7.07
## 10 F        22 Gold     310               3747      8.27
## # … with 79 more rows
```

```
## # A tibble: 100 x 6
## # Groups:   Sex, Age [100]
##    Sex     Age Medal  number total_silver_medal Percentage
##    <fct> <dbl> <fct>   <int>              <int>      <dbl>
##  1 F        11 Silver      1               3735     0.0268
##  2 F        12 Silver      3               3735     0.0803
##  3 F        13 Silver      6               3735     0.161
##  4 F        14 Silver     25               3735     0.669
##  5 F        15 Silver     59               3735     1.58
##  6 F        16 Silver    101               3735     2.70
##  7 F        17 Silver    100               3735     2.68
##  8 F        18 Silver    158               3735     4.23
##  9 F        19 Silver    179               3735     4.79
## 10 F        20 Silver    205               3735     5.49
## # … with 90 more rows
```

```
## # A tibble: 99 x 6
## # Groups:   Sex, Age [99]
##    Sex     Age Medal  number total_bronze_medal Percentage
##    <fct> <dbl> <fct>   <int>              <int>      <dbl>
##  1 F        12 Bronze      1               3771     0.0265
##  2 F        13 Bronze      2               3771     0.0530
##  3 F        14 Bronze     15               3771     0.398
##  4 F        15 Bronze     51               3771     1.35
##  5 F        16 Bronze     86               3771     2.28
##  6 F        17 Bronze    110               3771     2.92
##  7 F        18 Bronze    139               3771     3.69
##  8 F        19 Bronze    167               3771     4.43
##  9 F        20 Bronze    216               3771     5.73
## 10 F        21 Bronze    265               3771     7.03
## # … with 89 more rows
```

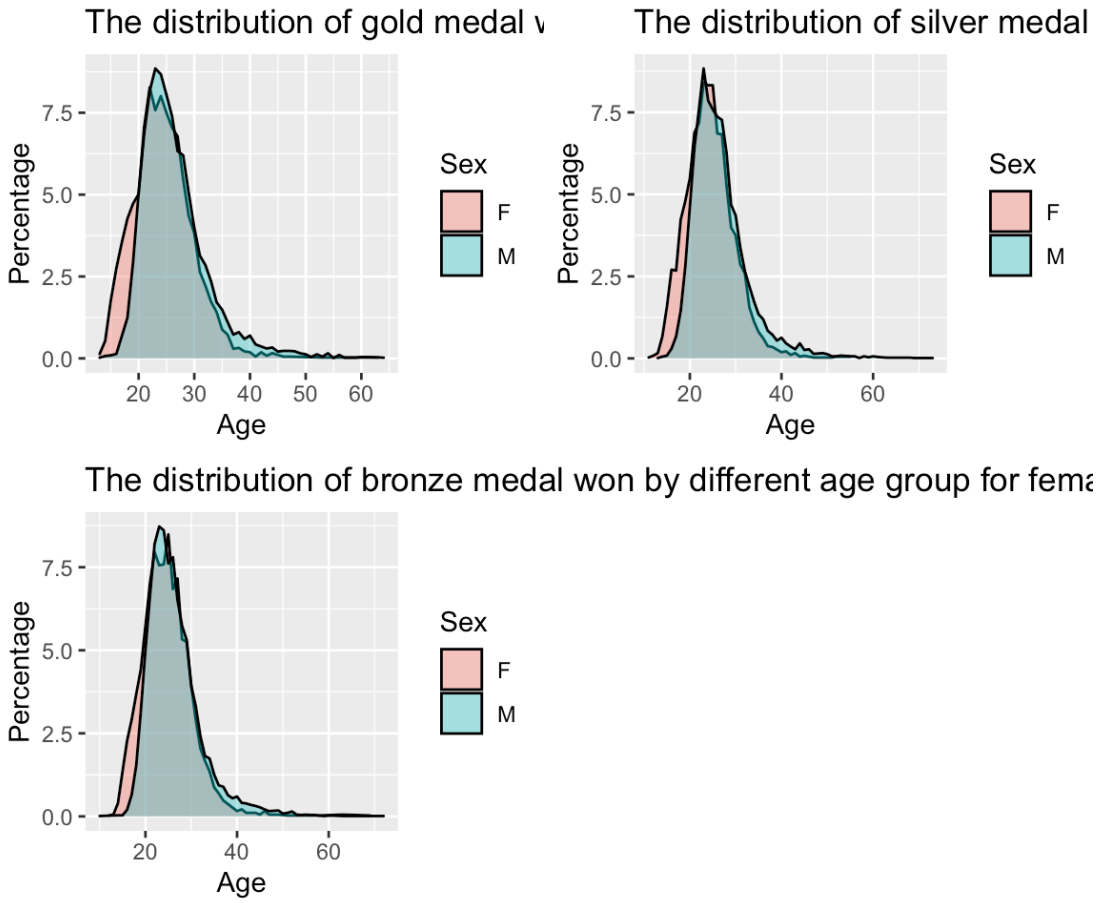

Figure 2.7: The medal distribution won in different age group by sex

n Figure 2.7, the different medal distribution was plotted for different age and compared in against both gender of male and female.

From this plot. the gold medal distribution for female and male are positively skewed with the athletes in the younger age group accounting for more of the gold medals earned than the older athletes (specifically, both female and male athletes age in early 20s have the highest percentage of the gold medal won).Similar result could be seen for silver medal distribution, both distribution for female and male athlete are positively skewed, however we can see that the age group that account the most percentage of silver medal won is ranged from 20-30 years older for both female and male athletes. In the bronze medal distribution, it shares similar distribution as silver medal as athletes age from 20 to 30 is account for the most percentage of bronze medal won for both female and male. Additionally, it is seen that for all medal distribution, female athletes of younger age (0-20) tend to account for higher percentage of medal (gold, silver, and bronze) than male athletes, however by age 30 and over, male athletes exceed female athletes in medal won for gold, silver and bronze medal. This could be due the quicker fall of physical, technical and strategics abilities of females athletes 30 and over , companied with increasing social pressure that female of an older age to be more family orientated.

Hence from the above analysis, we conclude that in general for both gender, age 20 to 30 tends to account for most proportions of the medals (gold, silver, and bronze)won, furthermore, it was found that female athletes tends to do better than male for medal won before age20, however later was exceed by male after age 30.

3 Conclusion

In summary, from the presentation today, we conclude that for quesiton1 when comparing the distribution of medal in different country aound the world, America appears to have the most medals in the world. In addtional, for question two, we have found that the sport athletics have the largest number of athletics, and in this sports, The United States accumulated the most gold MEDALS in each year. Furthermore for question three, we conclude that The United States

has an absolute advantage in swimming events,The medals of United Kingdom are more distributed in Cycling. and that the Medals from other countries are very evenly distributed among the five events lastly. when comparing the medals won for different age group by gender, it was revealed that in general for both gender, age bracket from 20-40 and 40 to 60 in general has the most proporiton of medal won compared to other age group for gold, silver and bronze medal won.

4 References

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686 (https://doi.org/10.21105/joss.01686)

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.6. https://CRAN.R-project.org/package=dplyr (https://CRAN.R-project.org/package=dplyr)

David Robinson, Alex Hayes and Simon Couch (2021). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.6. https://CRAN.R-project.org/package=broom (https://CRAN.R-project.org/package=broom)

Arel-Bundock et al., (2018). countrycode: An R package to convert country names and country codes. Journal of Open Source Software, 3(28), 848, https://doi.org/10.21105/joss.00848 (https://doi.org/10.21105/joss.00848)

Silge J, Robinson D (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *JOSS*, *1*(3). doi: 10.21105/joss.00037 (URL: https://doi.org/10.21105/joss.00037 (https://doi.org/10.21105/joss.00037)), <URL: http://dx.doi.org/10.21105/joss.00037 (http://dx.doi.org/10.21105/joss.00037)>.

Simon Garnier, Noam Ross, Robert Rudis, Antônio P. Camargo, Marco Sciaini, and Cédric Scherer (2021). Rvision - Colorblind-Friendly Color Maps for R. R package version 0.6.0.

South, Andy 2011 rworldmap: A New R package for Mapping Global Data. The R Journal Vol. 3/1 : 35-43.

D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. URL http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf (http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf)

Original S code by Richard A. Becker, Allan R. Wilks. R version by Ray Brownrigg. Enhancements by Thomas P Minka and Alex Deckmyn. (2018). maps: Draw Geographical Maps. R package version 3.3.0. https://CRAN.R-project.org/package=maps (https://CRAN.R-project.org/package=maps)

Pebesma, E.J., R.S. Bivand, 2005. Classes and methods for spatial data in R. R News 5 (2), https://cran.r-project.org/doc/Rnews/ (https://cran.r-project.org/doc/Rnews/).

Roger S. Bivand, Edzer Pebesma, Virgilio Gomez-Rubio, 2013. Applied spatial data analysis with R, Second edition. Springer, NY. https://asdar-book.org/ (https://asdar-book.org/)

Roger Bivand and Nicholas Lewin-Koh (2021). maptools: Tools for Handling Spatial Objects. R package version 1.1-1. https://CRAN.R-project.org/package=maptools (https://CRAN.R-project.org/package=maptools)

Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra (https://CRAN.R-project.org/package=gridExtra)

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ (https://www.R-project.org/).

Hadley Wickham and Dana Seidel (2020). scales: Scale Functions for Visualization. R package version 1.1.1. https://CRAN.R-project.org/package=scales (https://CRAN.R-project.org/package=scales)

Joe Cheng, Bhaskar Karambelkar and Yihui Xie (2021). leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library. R package version 2.0.4.1. https://CRAN.R-project.org/package=leaflet (https://CRAN.R-project.org/package=leaflet)

Hao Zhu (2021). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.4. https://CRAN.R-project.org/package=kableExtra (https://CRAN.R-project.org/package=kableExtra)

Bob Rudis (2020). hrbrthemes: Additional Themes, Theme Components and Utilities for 'ggplot2'. R package version 0.8.0. https://CRAN.R-project.org/package=hrbrthemes (https://CRAN.R-project.org/package=hrbrthemes)

Jason Cory Brunson and Quentin D. Read (2020). ggalluvial: Alluvial Plots in 'ggplot2'. R package version 0.12.3. http://corybrunson.github.io/ggalluvial/ (http://corybrunson.github.io/ggalluvial/)

Jeroen Ooms (2021). gifski: Highest Quality GIF Encoder. R package version 1.4.3-1. https://CRAN.R-project.org/package=gifski (https://CRAN.R-project.org/package=gifski)

Thomas Lin Pedersen and David Robinson (2020). gganimate: A Grammar of Animated Graphics. R package version 1.0.7. https://CRAN.R-project.org/package=gganimate (https://CRAN.R-project.org/package=gganimate)

Erich Neuwirth (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer (https://CRAN.R-project.org/package=RColorBrewer)

Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. https://igraph.org (https://igraph.org)

Thomas Lin Pedersen (2021). ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. R package version 2.0.5. https://CRAN.R-project.org/package=ggraph (https://CRAN.R-project.org/package=ggraph)

Hadley Wickham and Jim Hester (2020). readr: Read Rectangular Text Data. R package version 1.4.0. https://CRAN.R-project.org/package=readr (https://CRAN.R-project.org/package=readr)

Elmenshawy, A. R., Machin, D. R., & Tanaka, H. (2015). A rise in peak performance age in female athletes. Age, 37(3), 1-8.

Kaggle.com. 2021. 120 years of Olympic history: athletes and results. [online] Available at: https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results (https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results) [Accessed 23 May 2021].

Countries [Latitude & Longitude]. (2020, April 16). Kaggle. https://www.kaggle.com/franckepeixoto/countries (https://www.kaggle.com/franckepeixoto/countries)