Olympics Throughout History

A Deeper Dive into the data of the Olympics

Jackson Gasperack, Nina Mesyngier, Eric Farrall

2025-05-07

Table of contents

1 Introduction

The Olympics has been the golden standard for athletes of all sports for the last 130 years. To be an Olympic athlete means you must literally be one of the best in the world at your event. The Olympics was very intriguing to us because there is a wide variety of sports and countries and we thought there could be a lot of underlying relationships within the Olympics data over the years.

Image found from Chesnot/GettyImages (2024)

1.1 Primary Dataset

The data set we chose is a very large .csv file that we got from Kaggle. This file has more than 271,000 cases where a case represents an athlete that participated in an event in the given year. The data was recorded from the "First Olympics" in 1896 to the games in Rio de Janeiro in 2016.

The person who created the data set goes by the username rgriffin on Kaggle and, according to her Kaggle profile, she is the lead data scientist at Boston Consulting Group X. Most of the data was taken from the website www.sports-reference.com which is a central hub of decades of information about most modern day sports. This includes player statistics, physical attributes, and accolades for most players in any sport within the last 125 years.

The variables in this data set include:
- **ID**: Unique ID for each name
- **Name**: Name of the athlete
- **Sex**: M or F
- **Age**: Age of the athlete in that Olympics
- **Height**: Height of the athlete in that Olympics
- **Weight**: Weight of the athlete in that Olympics
- **Team**: The country the athlete competed for
- **NOC**: The country's three letter identifier
- **Games**: The Olympics Name (e.g. 2016 Summer)
- **Year**: The year the games took place
- **Season**: The time of year the games took place

- **City**: The city in which the games took place
- **Sport**: The sport the athlete competed in
- **Event**: The event the athlete competed in
- **Medal**: The type of medal the athlete received (N/A if none)

1.1.1 FAIR/CARE Principles

In terms of the FAIR principles, we believe that this data set embodies each quality. This data set was one of the first ones that appeared when searching up Olympics on Kaggle making it very findable and downloading it was easy, the only struggle was the size of the data set so we made a hugging face link for easier accessibility. Each person in our group has a different machine, yet the data set worked on each one making it interoperable. Despite the size of the file, it was easy to do data analysis and the author even included a detailed report she did herself for reference making this data easily reusable.

The CARE principles aren't as clear because their purposes focus more towards the creator of the data set than the actual data set itself. The principle of collective benefit is clear, because there is over 300 different data set notebooks on Kaggle that use this data set. We're not sure whether the creator has more permissions than us as people studying the data set, but she posted her wrangling/scraping code on GitHub and it seems to be read only unless you open the code in your own repository. The only principle that the creator might not have met is responsibility, because there are multiple little inconsistencies in the data set, according to the discussion, without any follow ups from the creator herself. Yet, this data set does seem to be ethical as she cited all of her sources and her code seems pretty unbiased.

1.2 Secondary Dataset

The other data set we chose also come from Kaggle and it is a .csv file of all the players that were drafted into the NBA from 1950-2018. We needed this data set to compare the names of this draft list of almost 70 years to the list of Olympic athletes. This data set also included .csv files of each player's position info and each player's season performance for a given year and team. Yet, those data sets were unnecessary to include because we only needed the names of the players.

The person who created this data set is Omri Goldstein who, according to his LinkedIn found on Kaggle, is a data scientist at Google. All of the data was scraped and wrangled from www.basketball-reference.com which is just a basketball focused version of the previously mentioned "sports-reference".

The variables in this data set include:
- **ID**: Unique identifier for each player
- **Name**: Name of the player
- **Height**: The player's height at the time of the draft
- **Weight**: The player's weight at the time of the draft
- **College**: The college the player was drafted from
- **Year of Birth**: The year the player was born
- **Birth city and state**: The city and state where the player was born

1.2.1 FAIR/CARE Principles

We believe that this data set also follows the principles of FAIR, as the data set was easily findable on Kaggle when we searched up the phrase NBA. This data was also easily accessible since the data set wasn't particularly large and it easily downloaded onto each group member's machine, making it interoperable. Finally, we thought this data was reusable because the .csv files which includes all player statistics for each season since 1950 has 53 different variables which easily allows for many different conclusions to be drawn.

We believe that this data set does not properly follow the principles of CARE. About 100 different data set notebooks were made on Kaggle from this data set and many of the discussion comments share their findings with others, which does follow the collective benefit principle. However, the creator of this data set did not share the code used to scrape and wrangle the data, which we believe is an abuse of authority to control giving the creator too many permissions compared to those studying. However, this creator did take responsibility as he is seen replying to most comments that bring up inconsistencies in the data set. We also believe that he followed the ethics principle as he cited his sources and got his data from an unbiased source.

1.3 Research Questions

When exploring the data in the data frame we found that the summer and winter Olympics were not always in different years. In fact, they used to be played in the same year up until 1992. We also found that the United States has a lot of gold medalists, and speaking of medalists the youngest one recorded in this data set was just 10 years old! We also saw history in the data by seeing West, East and unified Germany, along with the Soviet Union and Russia being two different country entries as well. These findings in our exploratory data analysis lead to us forming our research questions.

Research Questions:
1. What are the changes in the age distributions of athletes through the years?
2. How do various countries' participation in the Olympics change over time?
3. How does size (as measured in height and weight) affect the amount of medals that each athlete wins?

1.4 Download Necessary Packages

```
library(tidyverse)
library(ggplot2)
library(scales)
library(knitr)
library(kableExtra)
library(stringr)
library(patchwork)
```

You might need to install some of these packages if they aren't installed on your current machine already.

1.5 Load the Data into R

```
## Olympic Dataset
url <- "https://huggingface.co/datasets/EFarrallpsu/STAT184_Eric_Jackson_Nina/resolve/main/athl

athletes <- readr::read_csv(url)

## NBA Players R Dataset
url2 <- "https://gist.githubusercontent.com/jgasperackpsu/2c926a834c47eb5fc78d0626e7ff05bb/raw/

nba <- readr::read_csv(url2)
```

Both .csv files will be loaded through these links so everything should run properly regardless of if you have the files downloaded on your machine or not. The athletes data set was scraped and wrangled by rgriffin (2018), and the nba data set was scraped and wrangled by Goldstein (2018).

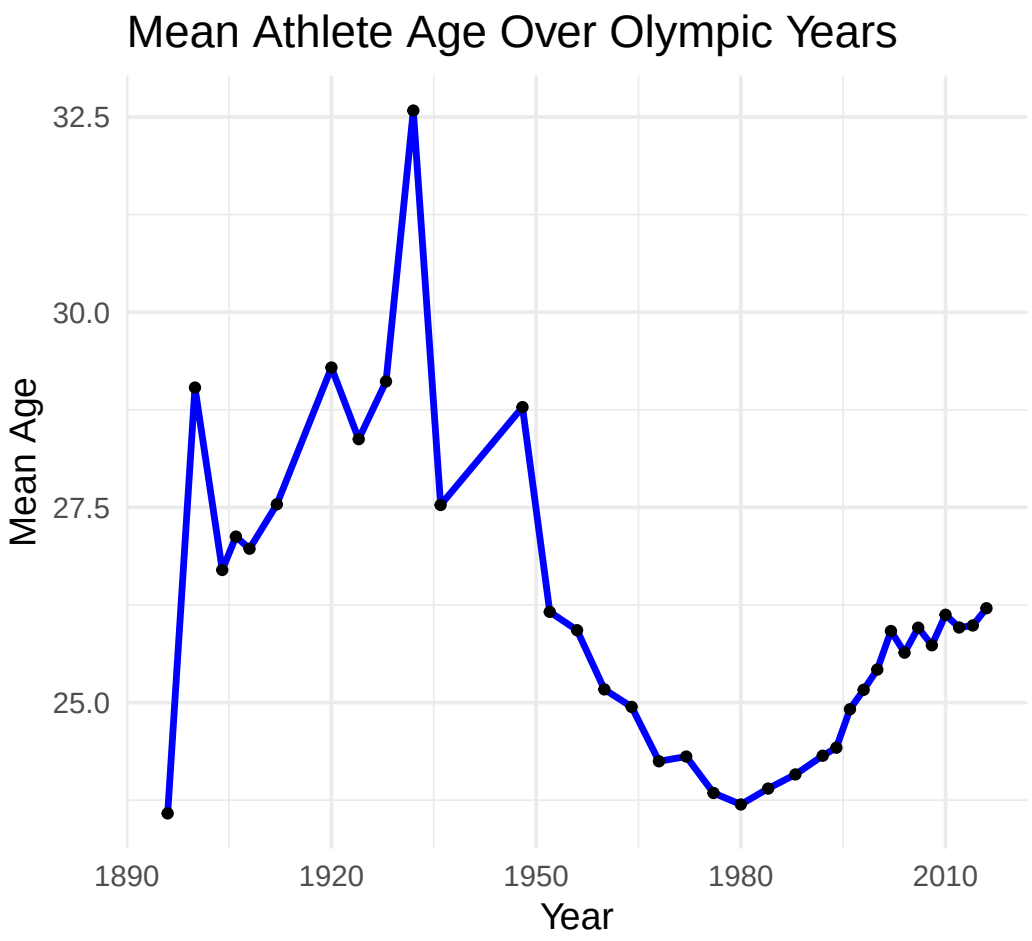
=======

2 Analysis on Age of Athletes

For the purpose of this section, the main attributes being focused on are Age and Year. The goal of this section is to explore if there are any trends in the frequency of the ages of athletes, or in the mean age of athletes.

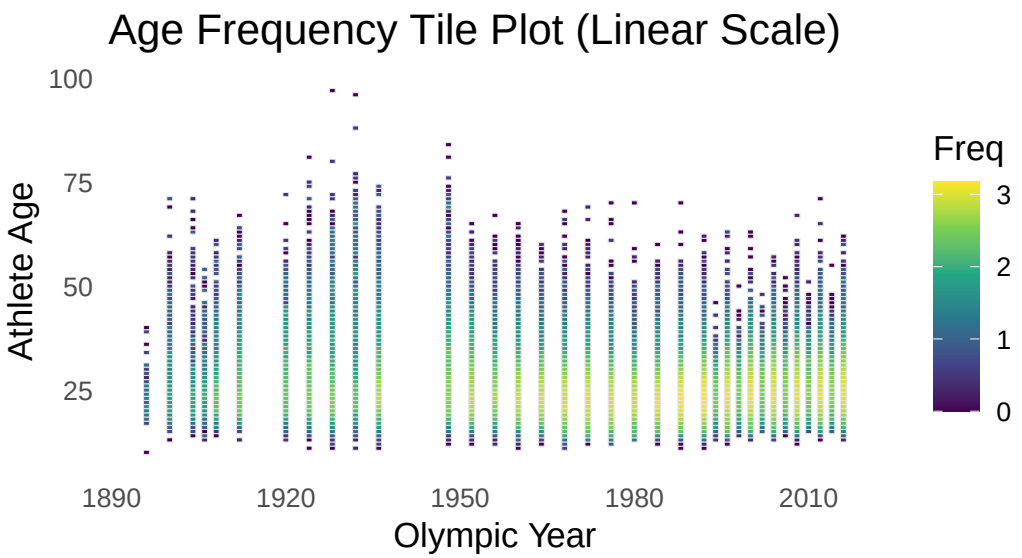
After cleaning the data up a bit, it is time to make a couple of visualizations to see if there are any obvious trends or outliers in the data set. The cleaning of the data can be seen in Section 5.2.1.

Figure 1: Mean Age of Olympic Athletes Over Olympic Years



This line graph shows the mean age of olympic athletes for a given year. As you can see, there was a big spike in the mean age between 1920 and 1950, as these were the years that Art Competitions were still a part of the olympics. We can also see that the lowest mean ages were around 1890 and 1980. Past 1980, the line increases until it begins to start leveling out jusr before 2010. The closer together data points occur because of the olympic games beginning to alternate which years were a Summer Olympics and which years were a Winter Olympics, meaning that there is less data per year, but more detailed data.

Figure 2: Frequency of Ages in Olympic Years



This figure shows the outliers that cause the mean age to spike around 1920. There are a couple of athletes in their late 90s who competed in Art Competitions. While further analysis could be done on the athletes who specifically competed in Art Competitions, there was a significant amount of age data missing from this subset of the cases, meaning whatever analysis was done, would not be fully accurate.

3 Analysis on Country-by-Country Olympic Participation Over Time

This section focuses on how countries' Olympic appearances change over time. An appearance is defined as an athlete participating in an event.

For our first graph, Figure 3, we have a coarse view of what was happening in terms of national participation. Coarse because the numerous gaps in the data make it a bit hard to see what is going on trend wise, but the gaps help show some significant Olympic history. This data focuses on only the top 8 countries with the greatest number of total appearances (as well as Russia and the Soviet Union) so the number of countries isn't overwhelming. Up until 1992, the summer and winter Olympic games happened in the same year, so the Olympics were every 4 years, not every 2 as it is now. Additionally, there's a significant period of no appearances between 1912-1920 and 1936-1948. This is due to the two world wars occurring in this time window. Germany and Japan didn't show up for additional year after WW2 due to be banned for apparent reasons. There is also a significant gap in German participation from 1968-1988, this happened because of Germany splitting into two parts, and registering under 2 different National Olympic Committee 3-digit codes, neither being 'GER'. The switch from Russia to Soviet Union, then from Soviet Union to Russia can be seen in the trade off in participation between URS and RUS at the bottom of Figure 3. However, there is a general trend of an increase of appearances over time for all countries, which can be seen more clearly in Figure 4, which has gaps removed for clarity. The process of wrangling these two graphs can be seen in Section 5.1.2.

Figure 3: Olympic Participation - Gaps

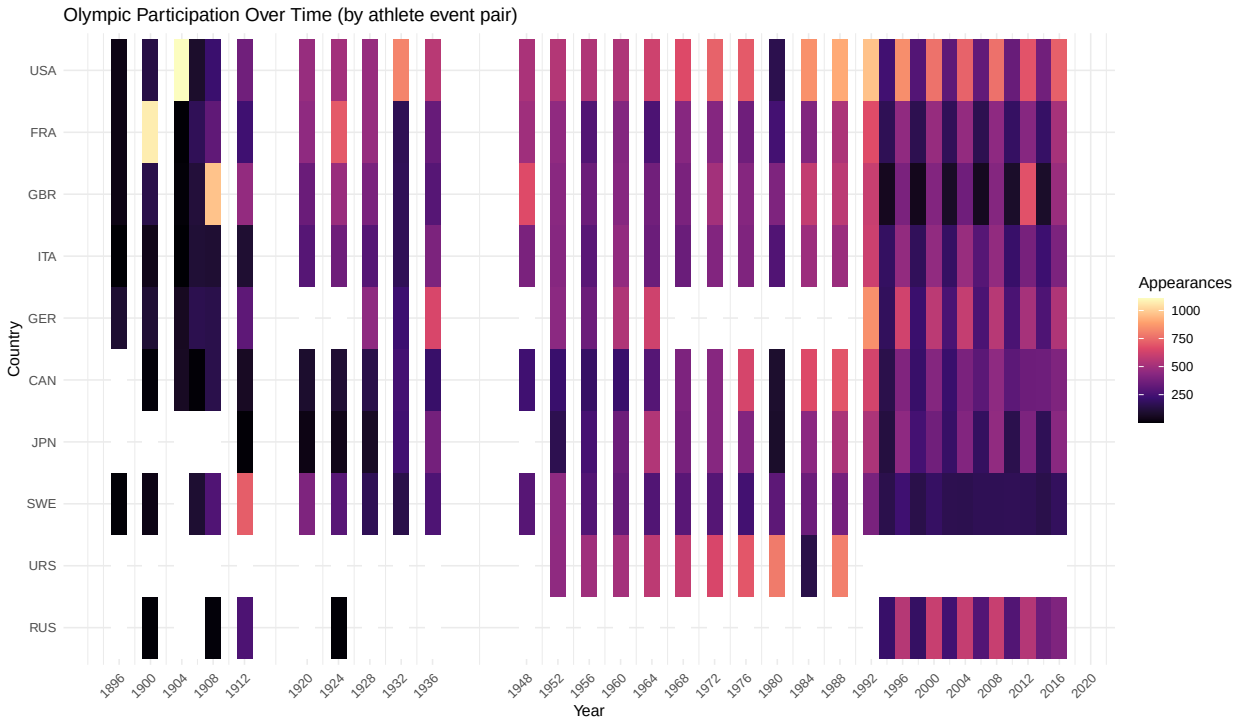
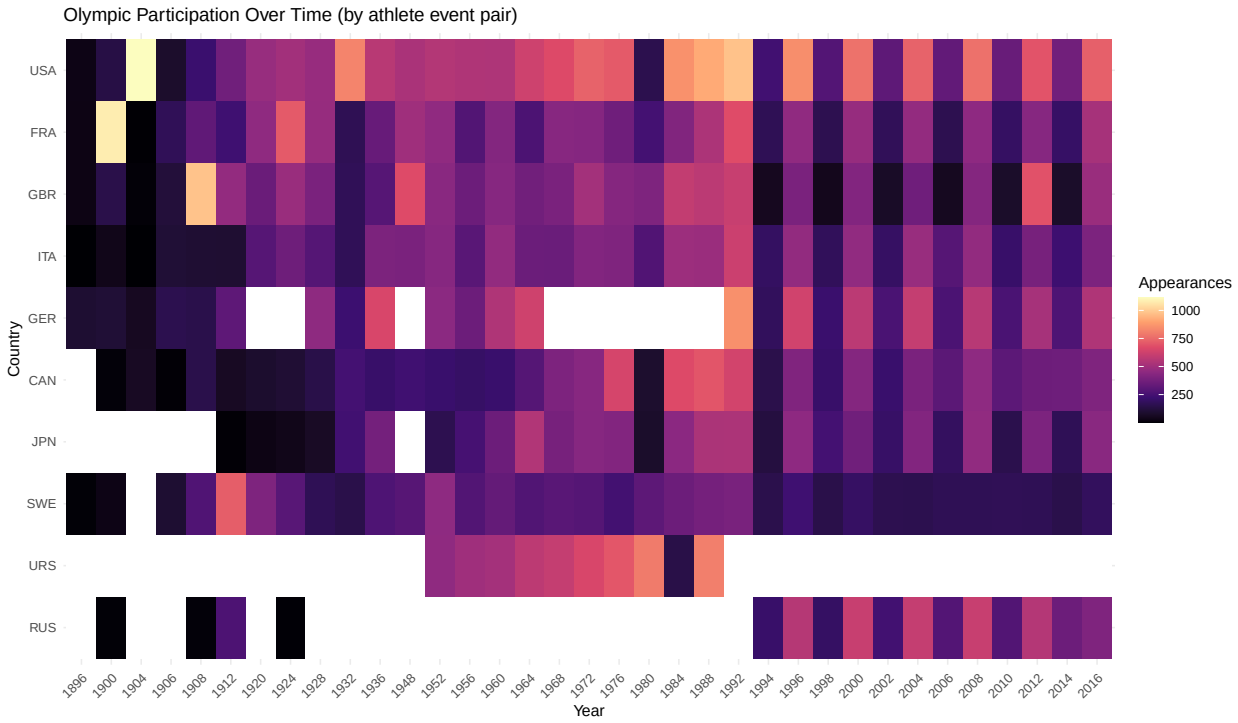


Figure 4: Olympic Participation - No Gaps



To better show trends, we made separate graphs for summer and winter games, as the switch from unified to split games in 1992 really throws off the interpretability of Figure 4, as summer Olympics generally have many more appearances than winter Olympics, mainly due to an increased number of available events to participate in. These graphs select which ten countries to show by total number of appearances for respectively the summer or winter games. The countries are ordered by most appearances at the top and least at the bottom (least still being 10th among all countries in the world). One can see that Canada (CAN) appears proportionally higher in the Winter games: it is ranked 2nd in Figure 6 and ranked 7th in Figure 5. Across all countries, it is apparent that appearances tend to increase over time. However, Hungary and Sweden seem to have an inflection point where their summer games participation starts to decrease. This could happen due to various reasons. Funding allocations may change within countries, whether due to specializing in certain sports or investing elsewhere in the country. Olympic qualification systems may change in a way that excludes smaller countries, or the culture of countries may change in a way hurtful to athleticism. Additionally, some countries definitely do specialize in a particular sport. This is apparent in the Canadian example, whose cold climate encourages them to pursue winter games more than summer games. The wrangling for the following two graphs can be seen in Section 5.1.3.

Figure 5: Summer Olympic Participation Over Time (by athlete event pair)

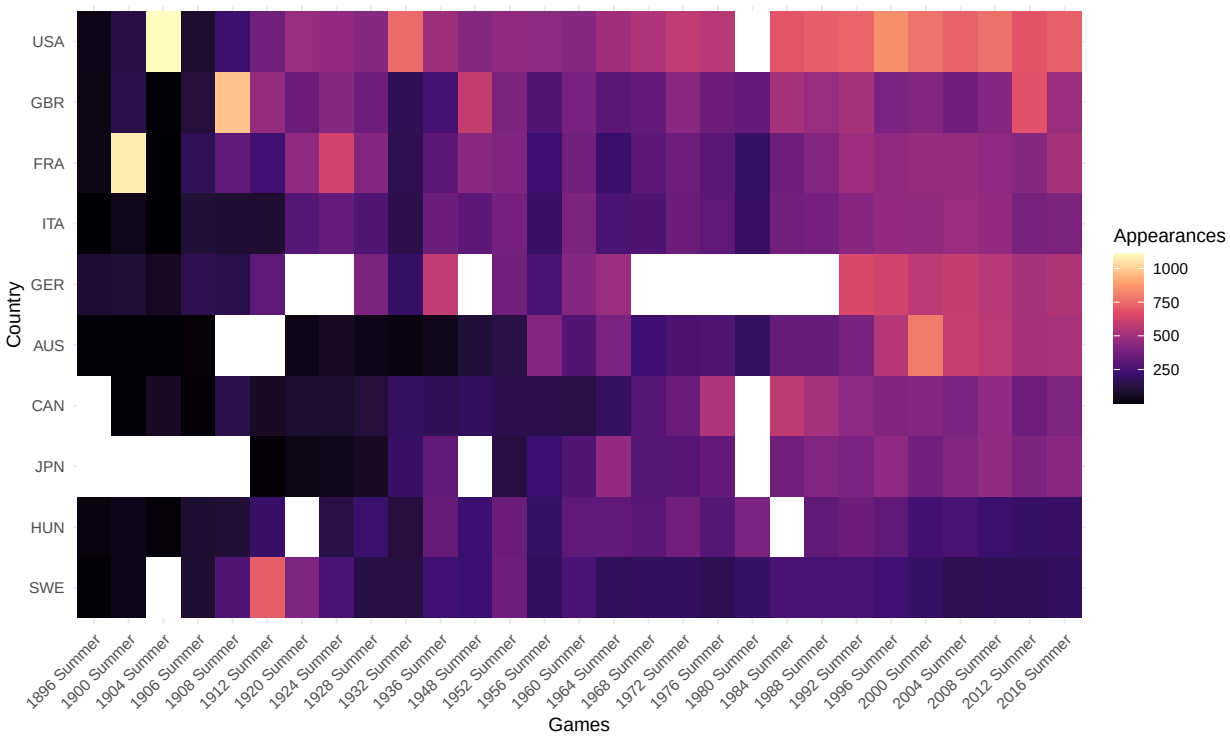
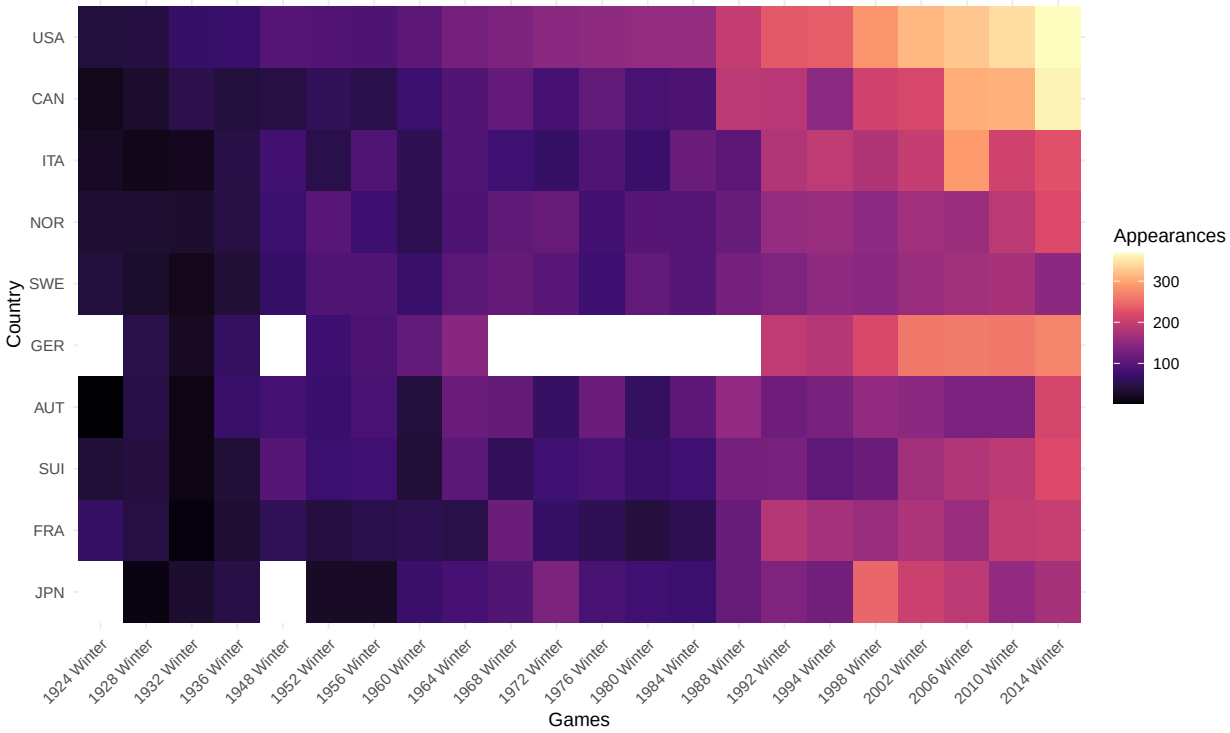


Figure 6: Winter Olympic Participation Over Time (by athlete event pair)



4 The Dominance of the United States

It is no secret that the United States is a heavy favorite to win the Olympic Games in total medal count every year. With fan favorite athletes like Michael Phelps, Simone Biles, and LeBron James, it's hard not to root for them. However, when we thought more about it, we started wondering what the U.S. was doing to be so far ahead of its competition.

We knew that it wasn't population, even though having more people in your country to choose from definitely helps, we are far better than countries like India and China in terms of total medals won. So, then we thought that size would be a good indicator of an athlete's success. This is where our research question about size predicting medals came from. We also wanted to focus on a singular sport to improve graph readability while also keeping in mind Kosslyn's Principle of Capacity Limitations. In our opinion, the sport where size, in terms of height, matters the most is basketball.

4.1 A Race for Second Place

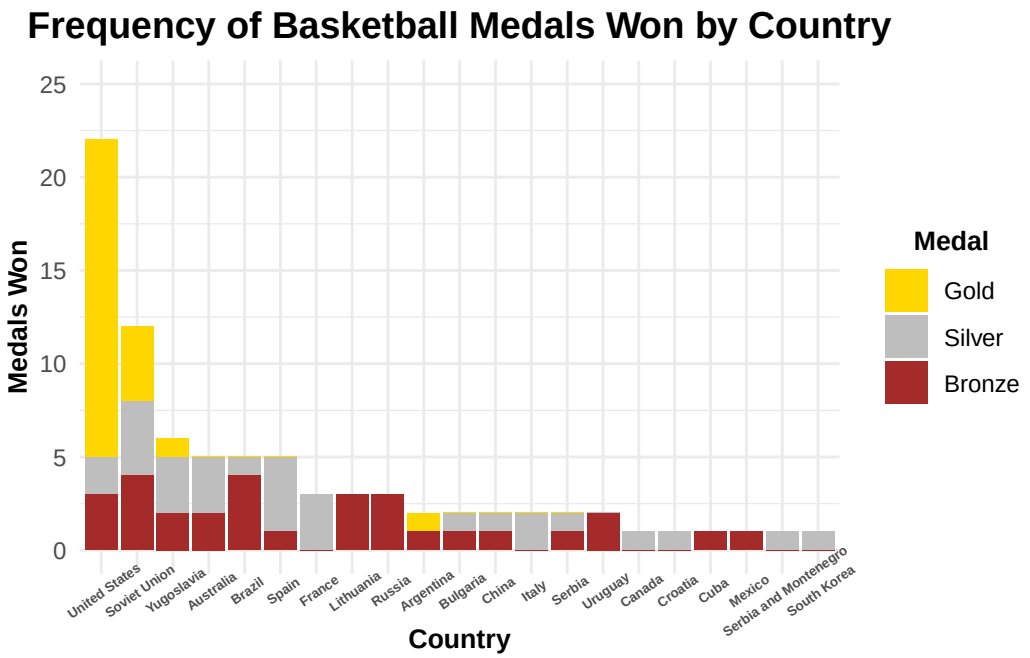


Figure 7: Country Basketball Medalists

Clearly, as shown above, the United States dominates when it comes to basketball. We have more gold medals alone than any other country has total medals. See Section 5.1.4 for the wrangling process used to make the data behind the data visualization. Seeing the sizable margin between first and second place made us think that the United States has overall bigger players than other countries. So, we decided to make a data visualization that showed how the height and weight of an athlete affects how many medals they've won. There's always a correlation between height and weight but we were looking for the group with the lowest medal count to be at the bottom of the line and it increases in medal count as the line goes up. So, we used our male and female basketball tables to wrangle medals per athlete and clean the data, which is all shown in Section 5.1.5.

4.2 The Effect of Size

Figure 8: Basketball Male Height vs Weight; Grouped by Medal Count

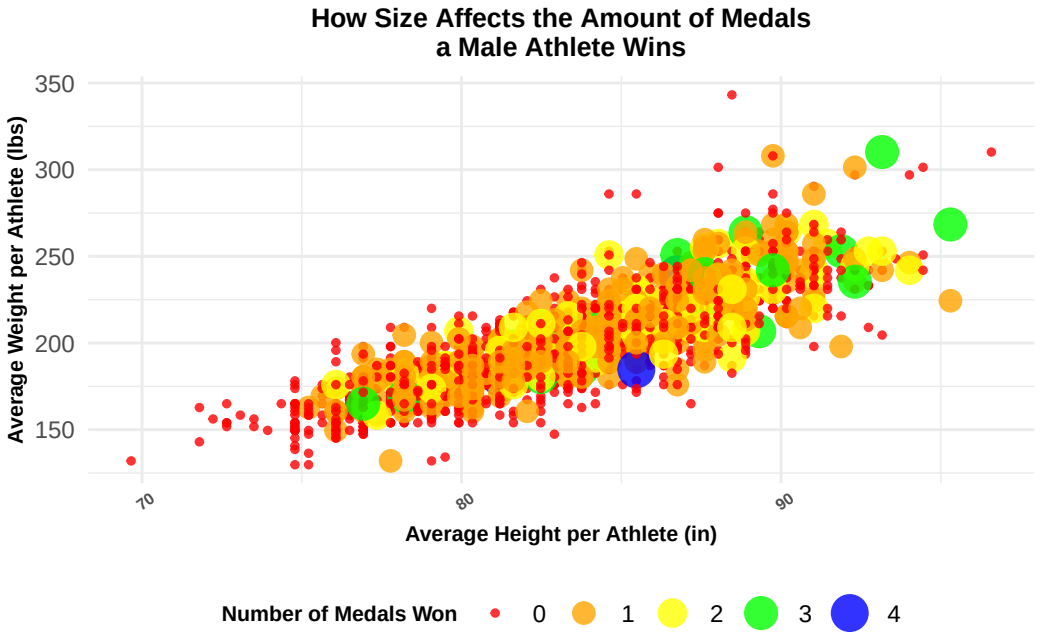
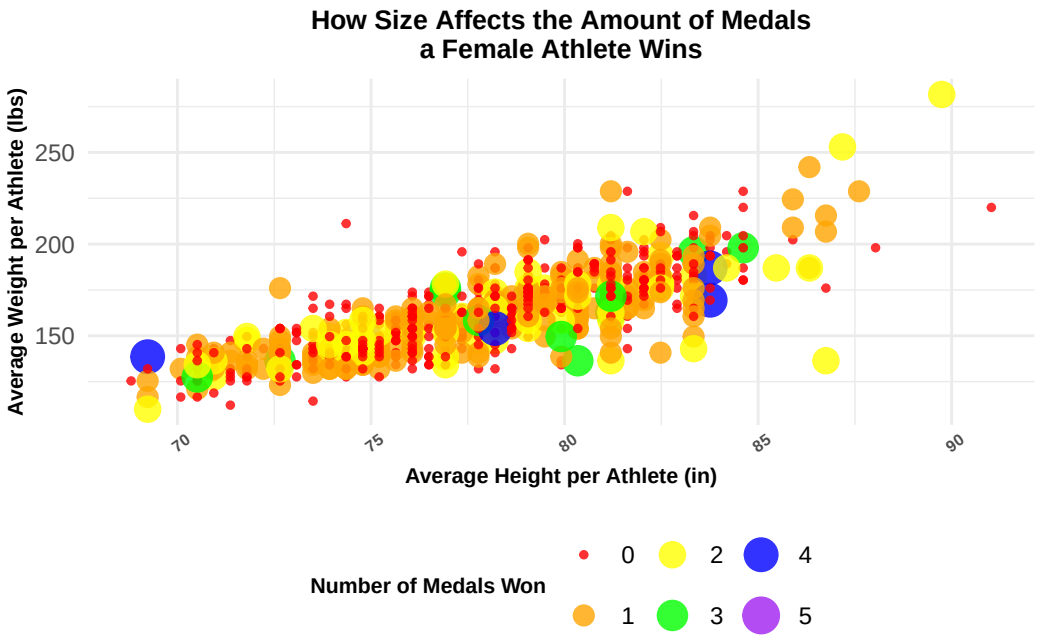


Figure 9: Basketball Female Height vs Weight; Grouped by Medal Count



Looking at both Figure 8 and Figure 9, we can see that there is no relationship between an athlete's height and weight and how many medals they won. For example, 7x WNBL all-star and 4x Olympic medalist, Kristi Harrower, is the blue dot near the bottom of the line in Figure 9 (BasketballAustralia (2017)). When we saw this data visualization, we were stumped because this was one of our main research questions and our findings were not what we expected.

We came up with some other variables that might be the reasoning behind the United States being an Olympic powerhouse. One we thought of was culture, because basketball is so popular in the U.S. more people are training to be as good as possible from a younger age. This was not consistent however, because the United States is not just good at basketball, they are good at every sport. Then, we thought of resources, like coaching, lifting equipment, and recovery. Note that there may be a relationship between medals won and the amount of resources at the athletes' disposal. Data like that is hard to find though because not every country makes information like that available to the public, but then we thought about the skill of the athletes.

4.3 NBA Presence

There is not a single defined metric for determining skill, but we *can* use the world's largest basketball league with the most skilled players in the world as a good determinant of it. Yes, the NBA is majorly an American league, but there are scouts constantly looking for new talent internationally. With names like Luka Doncic, Nikola Jokic, and Victor Wembanyama, the league has become a lot less American-heavy within the past couple decades. Therefore, we used data from the NBA data set we loaded in earlier to match the athletes that have been in both the NBA and the Olympics.

Image found from Irving (2024)

This part of the data wrangling phase was probably the most difficult. The Olympic data set had the names in a format that included the middle names. On the other hand, the NBA names had only first and last names, but some names had an asterisk on the end of them. Then we had to make all the names lower case so the join would match cases correctly. Then, we did the same process to get both the country the athlete participated for and the total number of basketball medals that country won. The entire wrangling process is in Section 5.1.6. However, once we had gotten it all wrangled and ready for the data visualization, we made a big finding.

Table 1: Number of NBA Players that Have Played for Each Team

Country	Number of Medals	Number of NBA Players
United States	22	196
Soviet Union	12	75
Yugoslavia	6	62
Australia	5	99
Brazil	5	106
Spain	5	85
France	3	94
Lithuania	3	47
Russia	3	27
Argentina	2	61
Bulgaria	2	37
China	2	87
Italy	2	99
Serbia	2	12
Uruguay	2	60
Canada	1	97
Croatia	1	32
Cuba	1	44
Mexico	1	64
Serbia and Montenegro	1	24
South Korea	1	55

Note:
Players who've played in multiple Olympics are counted once

We found that the amount of NBA players that have played for the United States is much larger than any other country. While there is not a consistent relationship for the two variables, the country with a large lead in medals also has a large lead in talented players. It is important to note that all athletes in the Olympic Games are skilled because they are among the best players in their country at their sport. The data suggests however, that the United States has the most skilled of those skilled players at their disposal due of the NBA.

This was a great finding for our team because at first we weren't sure if we would be able to continue with the research question. Yet, this only shows why the United States is dominant at basketball, but we want to show why the United States is dominant at all sports. As previously stated though, there is not a single defined metric for determining skill, which made this next section very tricky.

4.4 Winning World Wide

Since there is no defined measurement for skill, that means we would have had to search up a data set every sports' professional league and compare it to our Olympic data set for over 60 sports. That would have taken way too long for the time that we had, so we decided to go another route. We thought that if we could get the point across that medal count ***is not*** determined by the athletes' size then we could infer that it must be from another variable. Then, since we proved that the other variable happened to be a skill gap in basketball we can assume that the United States has more skillful athletes overall. It is important to note however, that the skill gap may not be the only variable which makes the United States win medals more than other countries. We stated earlier that the resources at the athletes' disposal could be a confounding variable, along with the privilege of the United States. Of course, the U.S. has its areas that struggle financially, but being a first world country allows more people to focus on things like sports rather than trying to afford food or shelter.

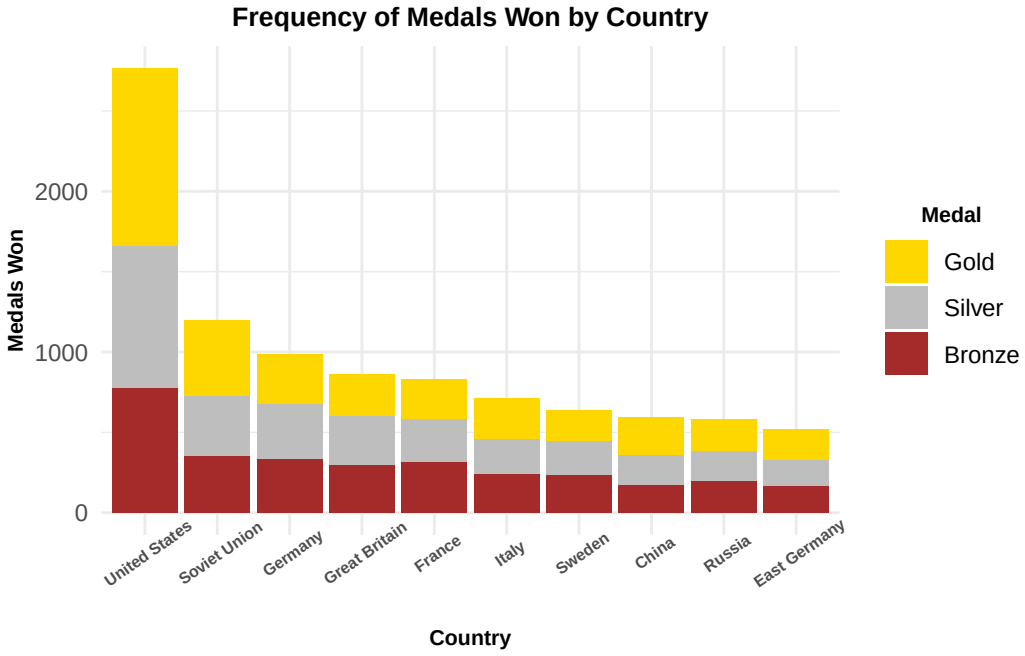


Figure 10: Medals Won by Each of the Top 10 Medal Winners

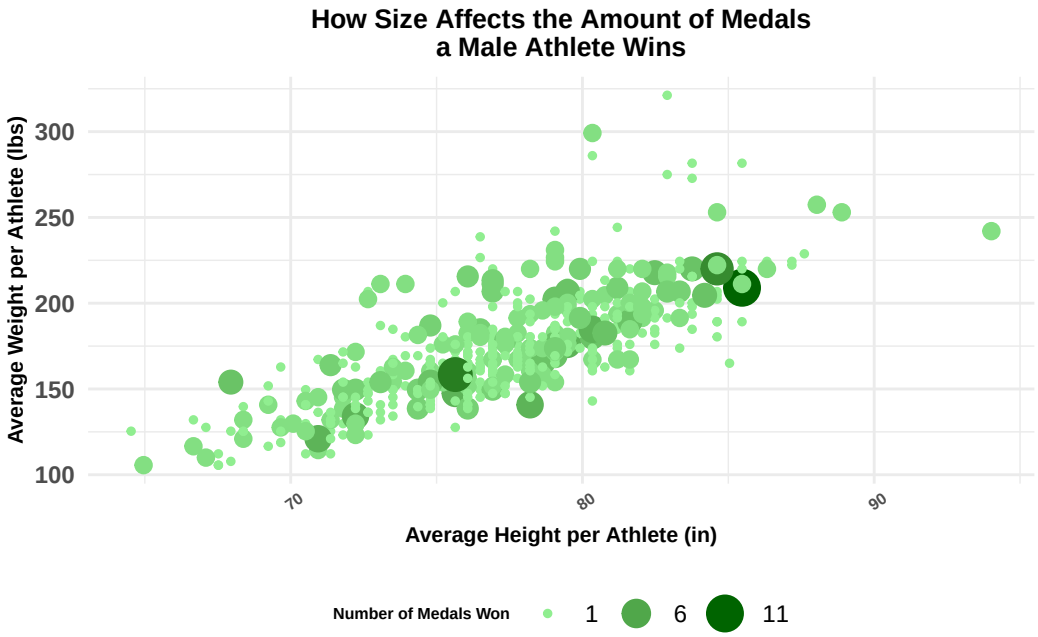
This visual really put it into perspective how dominant the United States is in the Olympic Games. This graph is slightly outdated but the gap has only widened since 2016 with the U.S. now having over 3000 total medals. The Soviet Union does not exist and has not since 1992 and they are still in second place, which honestly is very impressive. However, that means the United States almost has three times the amount of medals that the country in second place does, as of 2025 (Olympedia (2024)).

This visualization's data was hard to wrangle because we had to make sure team sports counted once towards a country's medal count. So, we made a table of all 700+ distinct events and looked for all keywords that indicated a team event. Then, we counted all individual event medals and team event medals separately and combined them to get the countries full medal count. Some countries also had duplicate entries odd endings and we had to remove the odd endings so all the

duplicates would form one. The rest of the wrangling had relatively the same process as Figure 7, and can be seen in Section 5.1.7.

The fact the United States has this far of a gap between them and second place indicates that there *must* be a variable that separates the United States' athletes from all of the other countries' athletes. Since we proved that size does not indicate an athlete's success in terms of medals won with basketball we decided to do the same with all sports. So, we did the same data wrangling process as the data behind Figure 8 and Figure 9 which can be seen in Section 5.1.8. The only difference is that this data set takes a random sample of 500 athletes from any sport.

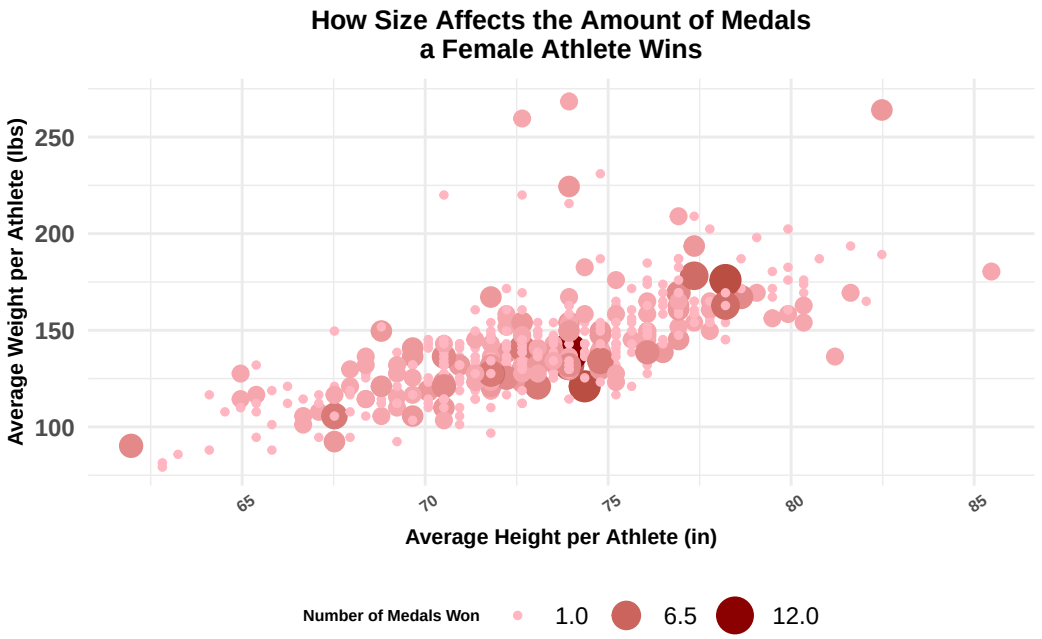
Figure 11: Male Height and Weight; Grouped by Medal Count



We actually cannot write a pre-written narrative text describing this data visualization because it changes with every render because it randomly samples 500 athletes from all sports each time. Sometimes it will show a relationship between size, in height and weight, and medal count, but in most cases it will show a plot with a positive correlation where dot size does not increase with it. Usually, dots of different sizes will be scattered all around the trend line. It is important to note that the legend is a scale from the left color to the right color with the minimum, midpoint, and maximum being the legend values.

This inconsistency in the different graphs' relationships allows us to rule out the graphs that have relationships as outliers. Since most of the time no relationship is shown we can assume that there is no relationship between an athlete's size and their medal count, regardless of the sports. Since neither population nor size are the variable that is the recipe for the United States' success, there must be another variable or combination of variables making the U.S. so dominant.

Figure 12: Female Height and Weight; Grouped by Medal Count



4.5 Analysis

We narrowed that the United States has several key variables that might be in their favor including:

- *Skill*: The U.S. athletes' are on average more skilled than other countries' athletes.
- *Diversity*: Having many different groups of people with different skill sets living in the same country definitely increases the number of people to choose from.
- *Privilege*: With the U.S. being a first world country a majority of children can put their focuses into sports without having to worry about what they are going to eat or where they are going to sleep. While, unfortunately, there are also children that live that way in the U.S., there are more children that live without worrying about that.
- *Resources*: The United States' athletes might have more and/or better resources than other countries because of the funding that sponsors and fundraisers bring to the team.

4.6 Conclusion

Through this analysis, we have a conclusion about what the United States has that separates them from the competition. We can not definitively say whether diversity plays a role in the States' Olympic success. We would have to find data for the diversity rating of each country and then compare that to the diversity of the entire Olympic team of each year to see if diversity increases medal count, which we did not have the time for. We also cannot say that having more resources causes the U.S. to win more medals because it's hard to keep track of something as broad as that. It would involve multiple data sets and a lot of missing information since not every country publishes

data like that. We also cannot prove that privilege plays an effect because, while first, second, and third world is a good broad term, it is not a good indicator for the privilege of the children in the country. Poverty rates would not work either because just because a child is not impoverished does not mean that they are privileged. So, it would be hard to know what to use as an indicator of "being privileged".

Therefore, that leaves us with one variable. With our analysis, the only variable where we can state that a relationship exists is the skill of the athletes. While the others could play a role in the amount of medals the U.S. wins, the only relationship that was presented in our data is that the gap in medals between the U.S. and the rest of the world is due to the gap in skill between U.S. athletes and other countries' athletes. This does not mean that every U.S. athlete is better than every other athlete in the world. Our data is just suggesting that the United States produces more skilled Olympic athletes compared to other countries.

5 Code Appendix

5.1 Wrangling

5.1.1 Age Analysis

```
# Remove NA ages and limit age range (optional for readability)
clean_data <- athletes %>%
  filter(!is.na(Age))

# Count frequencies by Age and Year
age_year_freq <- athletes %>%
  group_by(Year, Age) %>%
  summarise(Freq = log(n(),10), .groups = 'drop')

# Compute mean age per Olympic year
mean_age_by_year <- athletes %>%
  filter(!is.na(Age)) %>%
  group_by(Year) %>%
  summarise(Mean_Age = mean(Age), .groups = 'drop')
```

5.1.2 Wrangling Data to a Country-Year Case

```
#transform into data frame with case as year, with columns for each country
#signifying appearances
year_case <- athletes %>%
  pivot_wider(
    id_cols =  Year,
    names_from = NOC,
    values_from = NOC
  ) %>%
  arrange(Year)

# converts vector of strings signifying number of appearances to integer value
for (col in names(year_case)[-1]) {
  year_case[[col]] <- sapply(year_case[[col]],
                             function(x) if (is.null(x)) NA else length(x))
} #set null values as NA for better graphical differentiation later on

# changes case to country-year instead of country
country_year <- year_case %>%
  pivot_longer(cols = !Year, names_to = "Country", values_to = "Appearances")
```

```r
# identifies countries with most total appearances over the years
top_countries <- country_year %>%
  group_by(Country) %>%
  summarise(TotalAppearances = sum(Appearances, na.rm = TRUE)) %>%
  arrange(desc(TotalAppearances)) %>%
  slice_head(n = 8) %>%   # Get top 10
  pull(Country)

top_countries <- c(top_countries, "URS", "RUS") #because of russia politics

#filters to only include countries with top appearances
top_country_year <- country_year %>%
  filter(Country %in% top_countries)
top_country_year$Country <- factor(top_country_year$Country, levels = rev(top_countries))

# Olympic years (skipping canceled Olympics)
olympic_years <- c(1896, 1900, 1904, 1908, 1912, 1920, 1924, 1928, 1932, 1936,
                   1948, 1952, 1956, 1960, 1964, 1968, 1972, 1976, 1980, 1984,
                   1988, 1992, 1996, 2000, 2004, 2008, 2012, 2016, 2020)
```

5.1.3 Wrangling Data to Isolate for Seasons

```r
season_case <- athletes %>%
  pivot_wider(
    id_cols =  Games,
    names_from = NOC,
    values_from = NOC
  )

season_case <- season_case %>%
  mutate(Year = as.numeric(str_extract(Games, "^\\d{4}"))) %>%  # Extract year
  arrange(Year)                                                 # Sort by year

# converts vector of strings signifying number of appearances to integer value
for (col in names(season_case)[-1]) {
  season_case[[col]] <- sapply(season_case[[col]],
                               function(x) if (is.null(x)) NA else length(x))
}


# changes case to country-games instead of country
country_season <- season_case %>%
  pivot_longer(cols = !Games, names_to = "Country", values_to = "Appearances")
```

```r
country_summer <- country_season %>%
  filter(grepl("Summer", Games))

country_winter <- country_season %>%
  filter(grepl("Winter", Games))

# identifies countries with most total appearances over the years
top_countries_summer <- country_summer %>%
  group_by(Country) %>%
  summarise(TotalAppearances = sum(Appearances, na.rm = TRUE)) %>%
  arrange(desc(TotalAppearances)) %>%
  slice_head(n = 10) %>%   # Get top 10
  pull(Country)

top_countries_winter <- country_winter %>%
  group_by(Country) %>%
  summarise(TotalAppearances = sum(Appearances, na.rm = TRUE)) %>%
  arrange(desc(TotalAppearances)) %>%
  slice_head(n = 10) %>%   # Get top 10
  pull(Country)


#filters to only include countries with top appearances
country_summer <- country_summer %>%
  filter(Country %in% top_countries_summer)
country_summer$Country <- factor(country_summer$Country, levels = rev(top_countries_summer))

country_winter <- country_winter %>%
  filter(Country %in% top_countries_winter)
country_winter$Country <- factor(country_winter$Country, levels = rev(top_countries_winter))
```

5.1.4 Basketball Medals Per Country

```r
## All BB Athletes
basketball <- athletes %>%
  filter((`Sport` == "Basketball"))

## All Male BB Athletes
basketballMale <- basketball %>%
  filter(`Sex` == "M")

## All Female BB Athletes
basketballFemale <- basketball %>%
  filter(`Sex` == "F")
```

```
## Countries who've medaled in basketball
countryMedalists <- basketball %>%
  distinct(`Team`,`Year`,`Medal`) %>%
  group_by(`Team`) %>%
  summarise(
    goldWon = sum(`Medal` == "Gold", na.rm = TRUE),
    silverWon = sum(`Medal` == "Silver", na.rm = TRUE),
    bronzeWon = sum(`Medal` == "Bronze", na.rm = TRUE),
    total = sum(!is.na(`Medal`)),
    .groups = "drop"
  ) %>%
  filter(!(`Team` == "Unified Team")) %>% # Unified Team removed as it is not
# a real country but a combination of several
  filter(!(`total` == 0))

## Case is now Team and Medal type for Data Vis.
medalsPerBBTeam <- countryMedalists %>%
  pivot_longer(cols = c(goldWon,silverWon,bronzeWon),
               names_to = "Medal",
               values_to = "Count") %>%
  mutate(`Medal` = case_when(
    Medal == "goldWon" ~ "Gold", # Changing names of variables to medals
    Medal == "silverWon" ~ "Silver",
    Medal == "bronzeWon" ~ "Bronze"
  )) %>%
  mutate(Medal = factor(Medal, levels = c("Gold","Silver","Bronze")))
  # Changing order of medal types for better graph visuals
```

5.1.5 Basketball Medals Per Player

```
## All Male BB Athletes who won medals
basketballMedalsM <- basketballMale %>%
  group_by(`ID`,`Name`) %>%
  summarise(
    avgHeight = mean(Height, na.rm = TRUE),
    avgWeight = mean(Weight, na.rm = TRUE),
    goldWon = sum(`Medal` == "Gold", na.rm = TRUE),
    silverWon = sum(`Medal` == "Silver", na.rm = TRUE),
    bronzeWon = sum(`Medal` == "Bronze", na.rm = TRUE),
    total = sum(!is.na(`Medal`)),
    .groups = "drop"
  )

## All Female BB Athletes who won medals
basketballMedalsF <- basketballFemale %>%
```

```
  group_by(`ID`,`Name`) %>%
  summarise(
    avgHeight = mean(Height, na.rm = TRUE),
    avgWeight = mean(Weight, na.rm = TRUE),
    goldWon = sum(`Medal` == "Gold", na.rm = TRUE),
    silverWon = sum(`Medal` == "Silver", na.rm = TRUE),
    bronzeWon = sum(`Medal` == "Bronze", na.rm = TRUE),
    total = sum(!is.na(`Medal`)),
    .groups = "drop"
  )

## Removing all cases of NA in height and weight columns
## Converting to inches and pounds
maleBB_HandW <- basketballMedalsM %>%
  filter(!is.na(avgHeight)) %>%
  filter(!is.na(avgWeight)) %>%
  mutate(avgHeight = avgHeight/2.34,
         avgWeight = avgWeight*2.2)

femaleBB_HandW <- basketballMedalsF %>%
  filter(!is.na(avgHeight)) %>%
  filter(!is.na(avgWeight)) %>%
  mutate(avgHeight = avgHeight/2.34,
         avgWeight = avgWeight*2.2)
```

5.1.6 Joining NBA and Olympic Datasets

```
## Function to get only first and last names
getFirstLast <- function(x) {
  if (length(x) >= 2) {
    paste(x[1], x[length(x)])
  } else {paste(x[1])}
}

## Olympic Names
olympicBBNames <- basketballMale %>%
  select(Name) %>%
  mutate(Name = str_to_lower(str_squish(Name))) %>%
  mutate(first_last = sapply(str_split(Name, " "), getFirstLast)) %>%
# Make column without middle names to join two datasets
  rename("Player" = first_last)

## NBA Names
nbaNames <- nba %>%
  select(Player) %>%
```

```
  mutate(Player = str_to_lower(str_squish(Player))) %>%
  mutate(Player = str_replace_all(Player,"\\*$","")) # Remove asterisks

## Join
olympicNBA <- inner_join(olympicBBNames, nbaNames,
                         relationship = "many-to-many")

## Joining Countries
countryNameClean <- basketballMale %>%
  select(Name,Team) %>%
  mutate(Name = str_to_lower(str_squish(Name))) %>%
  mutate(first_last = sapply(str_split(Name, " "), getFirstLast)) %>%
  rename("Player" = first_last) %>%
  distinct(Player, .keep_all = TRUE)

olympicNBACountries <- right_join(olympicNBA, countryNameClean)

## Joining Medals
medalsNBACountries <- left_join(countryMedalists, olympicNBACountries,
                                by = "Team")

nbaPerCountry <- medalsNBACountries %>%
  distinct(Player, Team, .keep_all = TRUE) %>%
  group_by(Team) %>%
  summarise(
    Medals = first(total),
    numPlayers = n()
  ) %>%
  arrange(desc(Medals))
```

5.1.7 Total Medals Per Country

```
## Unique events
events <- athletes %>%
  select(Sport, Event) %>%
  distinct(Sport,Event) %>%
  select(-c("Sport")) %>% arrange(Event)

## Team Event Indicators
teamEvents <- c(
  "Team","Volleyball","Bobsleigh","Doubles","Baseball",
  "Fours","Relay","Basketball","Tandem","Pairs",
  "Handball","Hockey","Lacrosse","Polo","Group",
  "17-man","6-man","Eights","Quadruple","Rugby",
  "Two Person","Three Person","Softball","Tug-Of-War", "Football"
```

```
)

athletes <- athletes %>%
  mutate(
    isTeamEvent = str_detect(Event, str_c(teamEvents, collapse = "|"))
  )

## Making team sports have one combined medal
teamMedals <- athletes %>%
  filter(isTeamEvent, !is.na(Medal)) %>%
  distinct(Team, Games, Event, Medal)

indivMedals <- athletes %>%
  filter(!isTeamEvent, !is.na(Medal)) %>%
  distinct(Team, Games, Event, Medal)

## Combining team medals and individual medals
all_medals = bind_rows(teamMedals, indivMedals) %>%
  mutate(Team = str_replace_all(Team,"\\-1$","")) %>%  # Removing odd endings to
  mutate(Team = str_replace_all(Team,"\\-2$","")) %>%  # country names
  mutate(Team = str_replace_all(Team,"\\-3$","")) %>%
  mutate(Team = str_replace_all(Team,"\\-4$","")) %>%
  mutate(Team = str_replace_all(Team,"\\-15$",""))

## Medal counts
medalCountries <- all_medals %>%
  group_by(Team, Medal) %>%
  summarise(count = n(), .groups = "drop") %>%
  pivot_wider(
    names_from = Medal,
    values_from = count,
    values_fill = 0
  ) %>%
  mutate(total = Bronze + Gold + Silver)

## Case is now Team and Medal type for Data Vis.
medalsPerTeam <- all_medals %>%
  group_by(Team) %>%
  summarise(
    goldWon = sum(`Medal` == "Gold", na.rm = TRUE),
    silverWon = sum(`Medal` == "Silver", na.rm = TRUE),
    bronzeWon = sum(`Medal` == "Bronze", na.rm = TRUE),
    .groups = "drop"
  ) %>%
  pivot_longer(cols = c(goldWon, silverWon, bronzeWon),
               names_to = "Medal",
               values_to = "Count") %>%
```

```
  mutate(`Medal` = case_when(
    Medal == "goldWon" ~ "Gold", # Changing names of variables to medals
    Medal == "silverWon" ~ "Silver",
    Medal == "bronzeWon" ~ "Bronze"
  )) %>%
  mutate(Medal = factor(Medal, levels = c("Gold","Silver","Bronze"))) %>%
  group_by(Team) %>%
  mutate(total = sum(Count, na.rm = TRUE)) %>%
  ungroup() %>%
  arrange(desc(total)) %>%
  slice_head(n=30) # Top 10 teams (need 30, three medal types per country)
```

5.1.8 Total Medals Per Athlete

```
## Sample 500 random male athletes size and medal count for data vis.
heightWeightSampleM <- athletes %>%
  filter(Sex == "M") %>%
  group_by(ID, Name) %>%
  summarise(
    avgHeight = mean(Height, na.rm = TRUE),
    avgWeight = mean(Weight, na.rm = TRUE),
    goldWon = sum(`Medal` == "Gold", na.rm = TRUE),
    silverWon = sum(`Medal` == "Silver", na.rm = TRUE),
    bronzeWon = sum(`Medal` == "Bronze", na.rm = TRUE),
    total = sum(!is.na(`Medal`)),
    .groups = "drop"
  ) %>%
  filter(!is.na(avgHeight)) %>% filter(!is.na(avgWeight)) %>%
  filter(total > 0) %>%
  slice_sample(n = 500) %>%
  mutate(avgHeight_IN = avgHeight/2.34,
         avgWeight_LB = avgWeight*2.2)

## Sample 500 females
heightWeightSampleF <- athletes %>%
  filter(Sex == "F") %>%
  group_by(ID, Name) %>%
  summarise(
    avgHeight = mean(Height, na.rm = TRUE),
    avgWeight = mean(Weight, na.rm = TRUE),
    goldWon = sum(`Medal` == "Gold", na.rm = TRUE),
    silverWon = sum(`Medal` == "Silver", na.rm = TRUE),
    bronzeWon = sum(`Medal` == "Bronze", na.rm = TRUE),
    total = sum(!is.na(`Medal`)),
    .groups = "drop"
```

```
  ) %>%
  filter(!is.na(avgHeight)) %>% filter(!is.na(avgWeight)) %>%
  filter(total > 0) %>%
  slice_sample(n = 500) %>%
  mutate(avgHeight_IN = avgHeight/2.34,
         avgWeight_LB = avgWeight*2.2)
```

5.2 Data Visualizations

5.2.1 Mean Age Over Time

```
# Plot using ggplot2
averageAges <- ggplot(mean_age_by_year, aes(x = Year, y = Mean_Age)) +
  geom_line(color = "blue", size = 1.2) +
  geom_point(color = "black") +
  labs(title = "Mean Athlete Age Over Olympic Years",
       x = "Year",
       y = "Mean Age") +
  theme_minimal(base_size = 14)
```

5.2.2 Olympic Participation Over Time

```
#plot showing total appearances over time in top 10 countries
top10appearances <- ggplot(
  data = top_country_year,
  mapping = aes(
    x = Year,
    y = Country,
    fill = Appearances,
  )) +
  geom_tile(color = "transparent") +
  scale_fill_viridis_c(option = "A", na.value = 'white') +
  scale_x_continuous(
    breaks = olympic_years
  ) +
  theme_minimal() +
  labs(title = "Olympic Participation Over Time (by athlete event pair)",
       x = "Year",
       y = "Country") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
#plot with no empty space
top10NOgaps <- ggplot(
  data = top_country_year,
```

```
  mapping = aes(
    x = factor(Year),
    y = Country,
    fill = Appearances,
  )) +
  geom_tile(color = "transparent") +
  scale_fill_viridis_c(option = "A", na.value = 'white') +
  theme_minimal() +
  labs(title = "Olympic Participation Over Time (by athlete event pair)",
       x = "Year",
       y = "Country") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

5.2.3 Summer vs. Winter

```
#plot for summer games
top10summer <- ggplot(
  data = country_summer,
  mapping = aes(
    x = factor(Games),
    y = Country,
    fill = Appearances,
  )) +
  geom_tile(color = "transparent") +
  scale_fill_viridis_c(option = "A", na.value = 'white') +
  theme_minimal() +
  labs(x = "Games",
       y = "Country") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

#plot for winter games
top10winter <- ggplot(
  data = country_winter,
  mapping = aes(
    x = factor(Games),
    y = Country,
    fill = Appearances,
  )) +
  geom_tile(color = "transparent") +
  scale_fill_viridis_c(option = "A", na.value = "white") +
  theme_minimal() +
  labs(x = "Games",
       y = "Country") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

5.2.4 All Medal Winners for Basketball

```
## Medals for each country
medalsBBTeams <- ggplot(
  data = medalsPerBBTeam,
  mapping = aes(x = reorder(`Team`,-Count), y = `Count`, fill = `Medal`)
  # Order the bars by count
  ) +
  geom_bar(stat = "identity", position = "stack") + # Stacking diff. groups
  labs(
    title = "Frequency of Basketball Medals Won by Country",
    x = "Country",
    y = "Medals Won"
  ) +
  scale_fill_manual(values = c("Gold" = "gold","Silver" = "gray","Bronze" = "brown")) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5, size = 14),
    axis.title = element_text(face = "bold", size = 10),
    axis.text.x = element_text(face = "bold", size = 5, angle = 35),
    legend.title = element_text(face = "bold", size = 10, hjust = 0.5),
    axis.title.x = element_text(vjust = 1, margin = margin(t = -20))
    # Closes gap between axis title and axis text
  ) +
  scale_y_continuous(limits = c(0,25))
```

5.2.5 How Height and Weight Affect Basketball Medal Winners

```
## How size affects medals won for athletes
maleBB <- ggplot(
  data = maleBB_HandW,
  mapping = aes(
    x = avgHeight,
    y = avgWeight,
    size = total,
    color = total
    )
) +
  geom_point(alpha = .8) +
  labs(
    title = "How Size Affects the Amount of Medals\na Male Athlete Wins",
    x = "Average Height per Athlete (in)",
    y = "Average Weight per Athlete (lbs)",
    size = "Number of Medals Won",
    color = "Number of Medals Won"
```

```
  ) +
  scale_color_gradientn(colors = c("red","orange","yellow","green","blue")) +
  guides(size = guide_legend(), color = guide_legend()) + # Combine guides
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5, size = 10),
    axis.title = element_text(face = "bold", size = 8),
    axis.text.x = element_text(face = "bold", size = 6, angle = 35),
    legend.title = element_text(face = "bold", size = 8, hjust = 0.5),
    legend.position = "bottom"
  )

femaleBB <- ggplot(
  data = femaleBB_HandW,
  mapping = aes(
    x = avgHeight,
    y = avgWeight,
    size = total,
    color = total
  )
) +
  geom_point(alpha = .8) +
  labs(
    title = "How Size Affects the Amount of Medals\na Female Athlete Wins",
    x = "Average Height per Athlete (in)",
    y = "Average Weight per Athlete (lbs)",
    size = "Number of Medals Won",
    color = "Number of Medals Won"
  ) +
  scale_color_gradientn(colors = c("red","orange","yellow","green","blue","purple")) +
  guides(size = guide_legend(), color = guide_legend()) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5, size = 10),
    axis.title = element_text(face = "bold", size = 8),
    axis.text.x = element_text(face = "bold", size = 6, angle = 35),
    legend.title = element_text(face = "bold", size = 8, hjust = 0.5),
    legend.position = "bottom"
  )
```

5.2.6 NBA Players in the Olympics

```
nbaTable <- nbaPerCountry %>%
  knitr::kable("latex",
        col.names = c("Country", "Number of Medals", "Number of NBA Players"),
```

```
        booktabs = TRUE,
        escape = FALSE) %>%
  footnote("Players who've played in multiple Olympics are counted once") %>%
  kable_styling(latex_options = c("striped","scale_down")) %>%
  row_spec(0, bold = TRUE, font_size = 10)
```

5.2.7 Top 10 Medal Winners

```
## Medal Vis.
allSportsMedals <- ggplot(
  medalsPerTeam,
  mapping = aes(
    x = reorder(Team, -total),
    y = Count,
    fill = Medal
  )
) +
  geom_bar(stat = "identity", position = "stack") +
  labs(
    title = "Frequency of Medals Won by Country",
    x = "Country",
    y = "Medals Won"
  ) +
  scale_fill_manual(values = c("Gold" = "gold","Silver" = "gray",
                               "Bronze" = "brown")) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5, size = 10),
    axis.title = element_text(face = "bold", size = 8),
    axis.text.x = element_text(face = "bold", size = 6, angle = 35),
    legend.title = element_text(face = "bold", size = 8, hjust = 0.5)
  )
```

5.2.8 How Height and Weight Affect Olympic Medal Winners

```
## Show athlete size isn't why the US wins so much
maleALL <- ggplot(
  heightWeightSampleM,
  mapping = aes(
    x = avgHeight_IN,
    y = avgWeight_LB,
    size = total,
    color = total
```

```
  )
) + geom_point() +
  labs(
    title = "How Size Affects the Amount of Medals\na Male Athlete Wins",
    x = "Average Height per Athlete (in)",
    y = "Average Weight per Athlete (lbs)",
    size = "Number of Medals Won",
    color = "Number of Medals Won"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5, size = 10),
    axis.title = element_text(face = "bold", size = 8),
    axis.text.x = element_text(face = "bold", size = 6, angle = 35),
    axis.text.y = element_text(face = "bold"),
    legend.title = element_text(face = "bold", size = 6, hjust = 0.5),
    legend.position = "bottom",
    legend.box = "vertical"
  ) +
  guides(color = guide_legend(title = "Number of Medals Won"),
         size = guide_legend(title = "Number of Medals Won")) +
  scale_color_gradient(low = "lightgreen", high = "darkgreen",
                       breaks = c(min(heightWeightSampleM$total),
    (max(heightWeightSampleM$total)+min(heightWeightSampleM$total))/2,
                                  max(heightWeightSampleM$total))) +
  scale_size_continuous(breaks = c(min(heightWeightSampleM$total),
    (max(heightWeightSampleM$total)+min(heightWeightSampleM$total))/2,
                                   max(heightWeightSampleM$total)))


femaleALL <- ggplot(
  heightWeightSampleF,
  mapping = aes(
    x = avgHeight_IN,
    y = avgWeight_LB,
    size = total,
    color = total
  )
) + geom_point() +
  labs(
    title = "How Size Affects the Amount of Medals\na Female Athlete Wins",
    x = "Average Height per Athlete (in)",
    y = "Average Weight per Athlete (lbs)",
    size = "Number of Medals Won",
    color = "Number of Medals Won"
  ) +
  theme_minimal() +
```

```
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5, size = 10),
    axis.title = element_text(face = "bold", size = 8),
    axis.text.x = element_text(face = "bold", size = 6, angle = 35),
    axis.text.y = element_text(face = "bold"),
    legend.title = element_text(face = "bold", size = 6, hjust = 0.5),
    legend.position = "bottom",
    legend.box = "vertical") +
  guides(color = guide_legend(title = "Number of Medals Won"),
         size = guide_legend(title = "Number of Medals Won")) +
  scale_color_gradient(low = "lightpink", high = "darkred",
                       breaks = c(min(heightWeightSampleF$total),
    (max(heightWeightSampleF$total)+min(heightWeightSampleF$total))/2,
                                  max(heightWeightSampleF$total))) +
  scale_size_continuous(breaks = c(min(heightWeightSampleF$total),
    (max(heightWeightSampleF$total)+min(heightWeightSampleF$total))/2,
                                   max(heightWeightSampleF$total)))
```

References

BasketballAustralia. (2017). *Kristi karrower OLY*. https://www.australia.basketball/players/3597609

Chesnot/GettyImages. (2024). *Your guide to the paris 2024 summer olympics*. https://www.self.com/story/guide-to-paris-summer-olympics-2024

Goldstein, O. (2018). *NBA players stats since 1950*. https://www.kaggle.com/datasets/drgilermo/nba-players-stats/data

Irving, K. (2024). *Inside USAs basketball team: A complete roster and more to know about 2024 olympics 'dream team' comparison*. https://www.sportingnews.com/us/nba/news/usa-basketball-team-roster-2024-olympics/881312f8b876365feaa4b15e

Olympedia. (2024). *Medals by country*. https://www.olympedia.org/statistics/medal/country

rgriffin. (2018). *120 years of olympic history: Athletes and results*. https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results