Final Project Report

Code

Sarah Farooq

Due: May 4, 2021

Set Up

Hide

```
# clean up workspace environment
rm(list = ls())

# all packages used for the assignment
library(mosaic)
library(tidyverse)
library(esquisse)
library(ggplot2)
library(dcData)
library(dplyr)
```

Load Data into Environment

Hide

```
AthleteEvents <- read.csv("/Users/sarahfarooq/Desktop/athlete_events.csv")
NOC_regions <- read.csv("/Users/sarahfarooq/Desktop/noc_regions.csv")
data("CountryData")

head(AthleteEvents)
```

	ID <int>	Name <chr>	Sex <chr>	A... <int>	Height <int>	Weight <dbl>	Team <chr>	N... <chr>	Games <chr>
1	1	A Dijiang	M	24	180	80	China	CHN	1992 Summer
2	2	A Lamusi	M	23	170	60	China	CHN	2012 Summer
3	3	Gunnar Nielsen Aaby	M	24	*NA*	*NA*	Denmark	DEN	1920 Summer
4	4	Edgar Lindenau Aabye	M	34	*NA*	*NA*	Denmark/Sweden	DEN	1900 Summer
5	5	Christine Jacoba Aaftink	F	21	185	82	Netherlands	NED	1988 Winter
6	5	Christine Jacoba Aaftink	F	21	185	82	Netherlands	NED	1988 Winter

6 rows | 1-10 of 15 columns

Hide

```
head(NOC_regions)
```

	NOC <chr>	region <chr>	notes <chr>
1	AFG	Afghanistan	
2	AHO	Curacao	Netherlands Antilles
3	ALB	Albania	
4	ALG	Algeria	
5	AND	Andorra	
6	ANG	Angola	

6 rows

Hide

```
head(CountryData)
```

	country <chr>	area <dbl>	pop <dbl>	growth <dbl>	birth <dbl>	death <dbl>	migr <dbl>	maternal <dbl>	infant <dbl>
1	Afghanistan	652230	31822848	2.29	38.84	14.12	-1.83	460	117.23
2	Akrotiri	123	15700	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
3	Albania	28748	3020209	0.30	12.73	6.47	-3.31	27	13.19
4	Algeria	2381741	38813722	1.88	23.99	4.31	-0.93	97	21.76
5	American Samoa	199	54517	-0.35	22.87	4.68	-21.64	*NA*	8.92
6	Andorra	468	85458	0.17	8.48	6.82	0.00	*NA*	3.69

6 rows | 1-10 of 76 columns

Research Question: As many countries experience rapid technological, political, and societal development over the years, which kind of factors impact how many medals a country wins in the Olympics?

Primary Data Set: Clean Up

- The primary data set for this research question is the AthleteEvents data.This data set contains information on the modern Olympic games. It includes data from Athens 1896 to Rio 2016. It is important to know that the Olympics host thousands of athletes from over 200 countries. They participate in a variety of competitions. In this data set, each row (case) represents an individual athlete competing in a certain Olympic event. There are 271116 rows. For each athlete, important information is provided through 15 key variables: ID, Name, Sex, Age, Height, Weight, Team, NOC, Games, Year, Season, City, Sport, Event, and Medal.
- In this data set, the “Medal” variable indicates which kind of medal the athlete won in their event (Gold, Silver, Bronze). If the athlete did not win a medal, there is a “NA,” or missing value, in this spot. These missing values can influence graphical models/ modeling decisions and cause confusion, so it would be better to change these NA values to “No Medal.” By having the value “No Medal,” we can perform a count of how many athletes do not win any medals, making the analysis of the research question more comprehensible. To do so:

Hide

```
AthleteEvents <- # table with appropriate medal values stored into new table
  AthleteEvents %>%
  replace_na(list(Medal = "No Medal"))

AthleteEvents %>%
  head(5)
```

	ID <int>	Name <chr>	Sex <chr>	A... <int>	Height <int>	Weight <dbl>	Team <chr>	N... <chr>	Games <chr>
1	1	A Dijiang	M	24	180	80	China	CHN	1992 Summer
2	2	A Lamusi	M	23	170	60	China	CHN	2012 Summer
3	3	Gunnar Nielsen Aaby	M	24	*NA*	*NA*	Denmark	DEN	1920 Summer
4	4	Edgar Lindenau Aabye	M	34	*NA*	*NA*	Denmark/Sweden	DEN	1900 Summer
5	5	Christine Jacoba Aaftink	F	21	185	82	Netherlands	NED	1988 Winter

5 rows | 1-10 of 15 columns

Second Data Set: Combine with Primary Data Set

- The second data set for this research question is the NOC_regions data set. This data set contains information on each NOC (National Olympic Committee 3 letter code). It provides the country name that corresponds with each NOC, as well as additional clarifying notes for each NOC (however, there are very few notes). There are only 3 key variables in this data set: NOC, region, notes.
- This data set is helpful because it can be joined with the AthleteEvents data to identify which country each athlete is from, instead of trying to decipher this information from just the NOC variable in the table. From the table below, we can see that from joining the two sets, the AthleteEvents data set has two additional columns called “region” and “notes.”

Hide

```
AthleteEvents <- # the joining of the two tables will be stored in the athletes table
  AthleteEvents %>%
  inner_join(NOC_regions, by = "NOC")

head(AthleteEvents)
```

	ID <int>	Name <chr>	Sex <chr>	A... <int>	Height <int>	Weight <dbl>	Team <chr>	N... <chr>	Games <chr>
1	1	A Dijiang	M	24	180	80	China	CHN	1992 Summer
2	2	A Lamusi	M	23	170	60	China	CHN	2012 Summer
3	3	Gunnar Nielsen Aaby	M	24	*NA*	*NA*	Denmark	DEN	1920 Summer
4	4	Edgar Lindenau Aabye	M	34	*NA*	*NA*	Denmark/Sweden	DEN	1900 Summer
5	5	Christine Jacoba Aaftink	F	21	185	82	Netherlands	NED	1988 Winter
6	5	Christine Jacoba Aaftink	F	21	185	82	Netherlands	NED	1988 Winter

6 rows | 1-10 of 17 columns

Third Data Set: Clean Up

- The third data set for this research question is the CountryData data set, found from the dcData package in RStudio. This data set is from the CIA Factbook, and it contains information on geographic, demographic, and economic data on a country-by-country basis. There are 76 variables on each of the 256 countries (cases) in the world. However, due to the scope of the research question, there are only a selected amount of variables that I believe are useful for this project. Therefore, only the following variables from the CountryData data set are considered for the analysis: country, pop, growth, life, health, GDP, GDPgrowth, GDPcapita, and debt. By only using these variables, the CountryData data set will look like:

Hide

```
CountryData <- # storing only necessary variables into CountryData table
  CountryData %>%
  select(country, pop, growth, life, health, GDP, GDPgrowth, GDPcapita, debt)

CountryData %>%
  head(5)
```

	country <chr>	pop <dbl>	growth <dbl>	life <dbl>	health <dbl>	GDP <dbl>	GDPgrowth <dbl>	GDPcapita <dbl>	debt <dbl>
1	Afghanistan	31822848	2.29	50.49	9.6	4.530e+10	3.1	1100	*NA*
2	Akrotiri	15700	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
3	Albania	3020209	0.30	77.96	6.3	2.834e+10	0.7	10700	70.5
4	Algeria	38813722	1.88	76.39	3.9	2.847e+11	3.1	7500	13.2
5	American Samoa	54517	-0.35	74.91	*NA*	5.753e+08	3.0	8000	*NA*

5 rows

Data Analysis

- The research question we are exploring are the factors that impact how many medals a country wins at the Olympics.

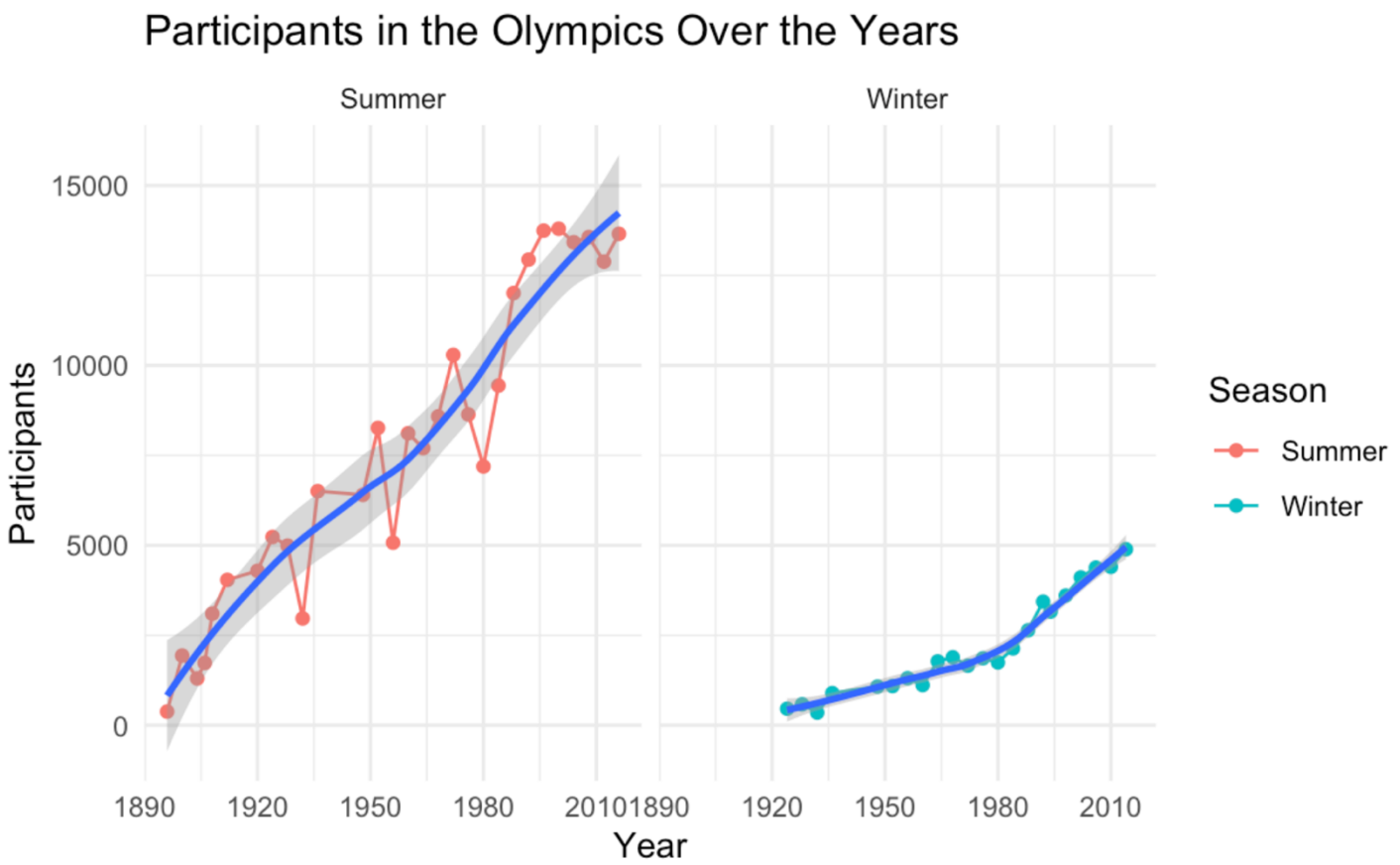
The Number of Participants Over the Years

- To begin, the first aspect about the Olympics which contributes to how many medals a country wins is the amount of participants in the games. It is most likely that as the Olympics and countries participating progressed over time, the amount of athletes participating increased. The following graphic shows the increase in participants according to each season.

Hide

```
AthleteEvents %>%
  group_by(Year, Season) %>%
  summarise(Participants = n()) %>%
  ggplot(aes(x = Year, y = Participants)) +
  facet_wrap(~ Season) +
  geom_point(aes(color = Season)) +
  geom_line(aes(color = Season)) +
  geom_smooth() + # smoother needed to show general trend of increasing participants
  theme_minimal() +
  labs(title = "Participants in the Olympics Over the Years")
```

```
`summarise()` has grouped output by 'Year'. You can override using the `.groups` argument.
```



- From the graph above, we see that participation in the Winter Olympics slowly increased over time. For the Summer Olympics, there were sharp increases and decreases during certain periods, but the overall trend showed that participation increased over time.
- These trends of increasing and decreasing participation could be results of countries possibly being in economic or political distress, or countries could be flourishing economically and progressing socially. This may lead countries to make financial decisions on how many athletes they can send to the Olympics to represent them. It is important to have this knowledge about participation trends to have some background about the Olympics, a better understanding of the data sets, and to perform further analysis.

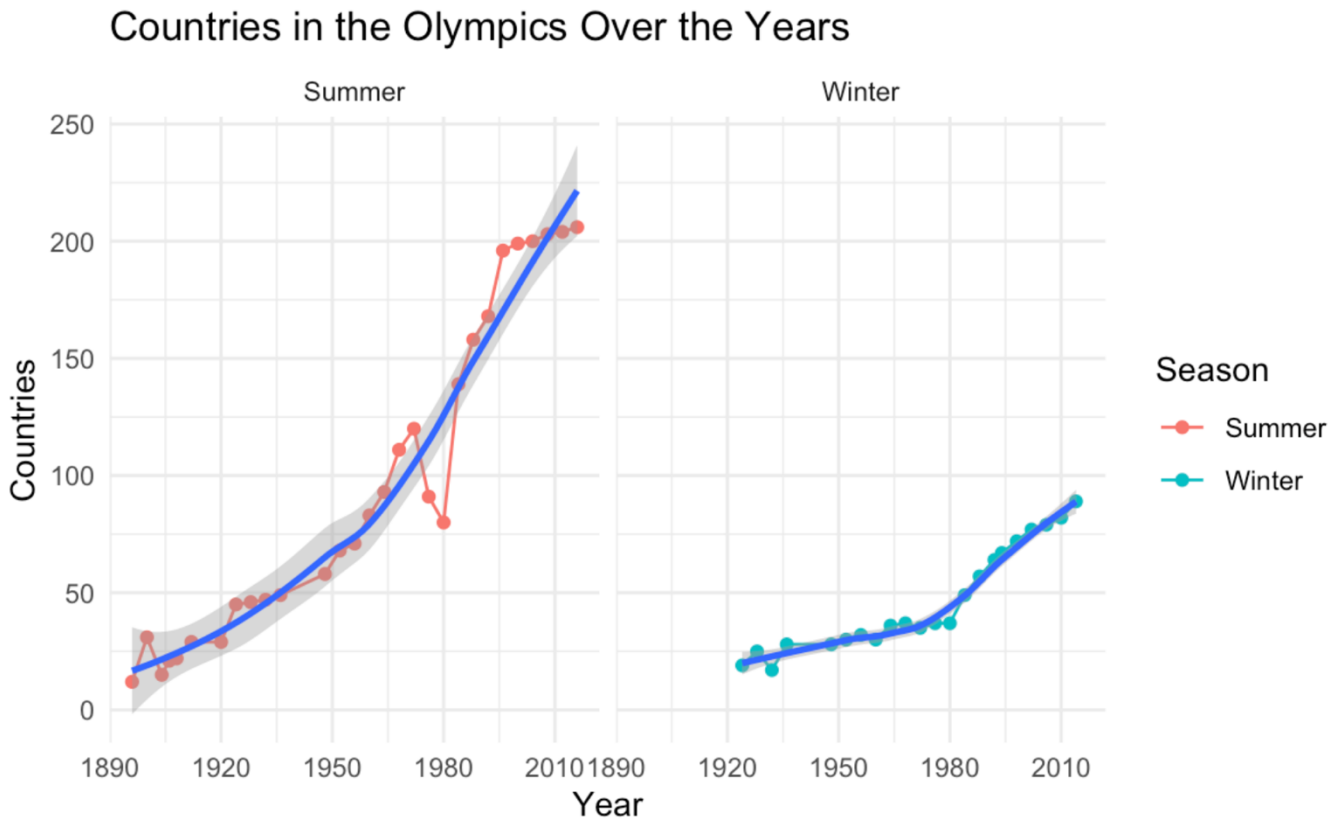
The Number of Countries Over the Years

- Another aspect about the Olympics to observe and be knowledgeable of is how many countries participated over the years. As countries progress technologically and politically, it is likely that the number of nations competing has increased over time. The following graphic shows the increase in countries according to each season.

Hide

```
AthleteEvents %>%
  group_by(Year, Season) %>%
  summarise(Countries = length(unique(NOC))) %>%
  ggplot(aes(x = Year, y = Countries)) +
  facet_wrap(~ Season) +
  geom_point(aes(color = Season)) +
  geom_line(aes(color = Season)) +
  geom_smooth() + # smoother needed to show general trend of increasing countries
  theme_minimal() +
  labs(title = "Countries in the Olympics Over the Years")
```

```
`summarise()` has grouped output by 'Year'. You can override using the `.groups` argument.
```



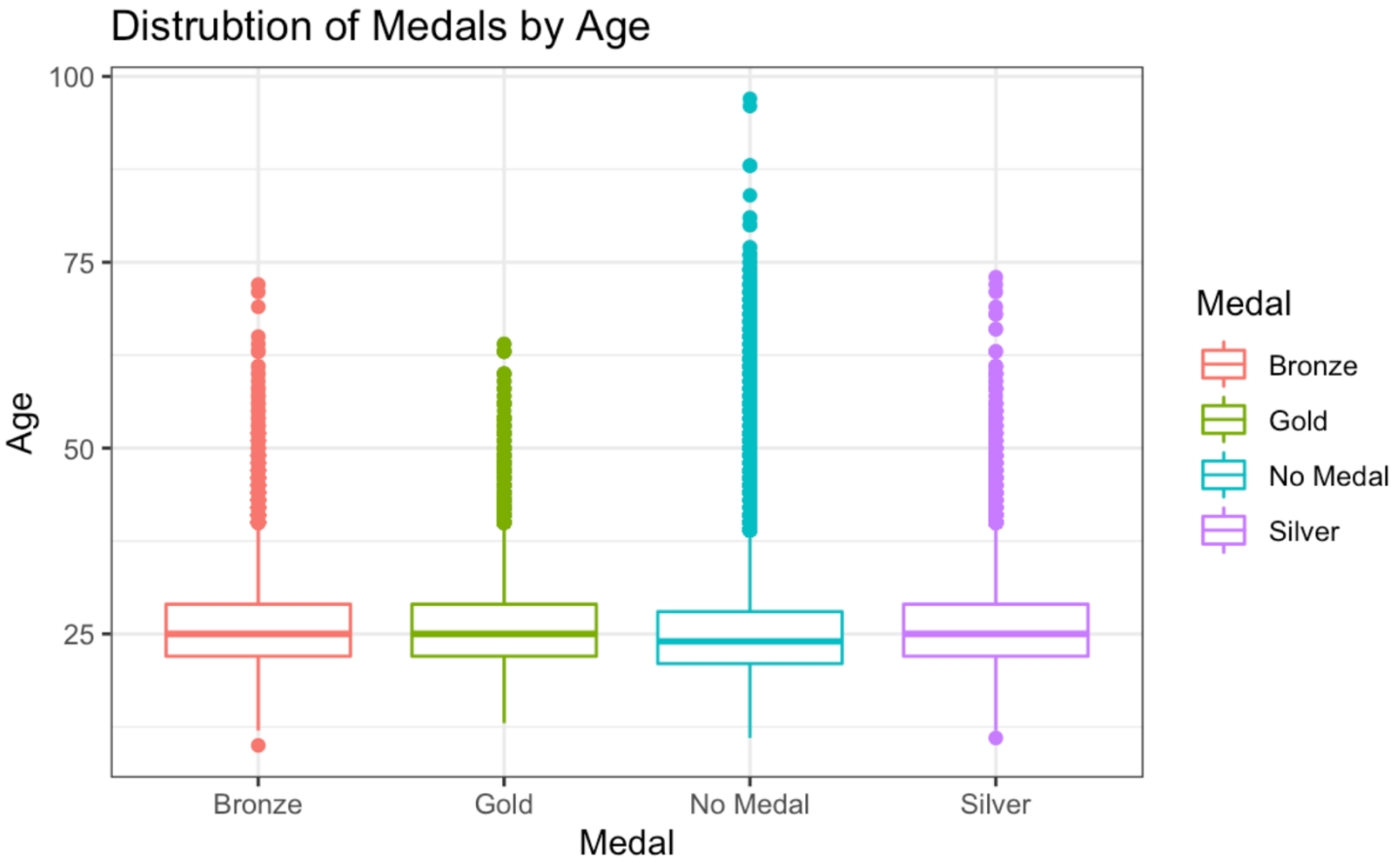
- Similar to the number of participants over time, the graph above shows that countries participating in the Winter Olympics slowly increased over time. The number of countries participating in the Summer Olympics faced some sharp increases and decreases as well, but show a general increase in country participation.
- These trends could once again be a result of countries facing different kinds of economic and political situations. For example, both graphs above (participants and countries over time) display a sharp decrease in participation for the Summer Olympics in the year 1980. This was a result of the 1980 Olympics boycott, started by the United States to protest the Soviet invasion of Afghanistan. In 1980, only 80 countries participated in the Summer Olympics, explaining the low number of athletes.
- As countries progress financially and technologically, they likely have more resources to train athletes and send them to the Olympics to represent their country. The sharp increase can be seen prominently in the more recent years. It is interesting to observe these trends and be knowledgeable of them for further analysis.

The Correlation Between Medals and Age

- We will be exploring how different factors such as the life expectancy and general health of citizens from specific countries impacts the amount of medals won. Therefore, another factor to observe is the age of each participant. It is possible that the age of the participant influences which medal they win (Gold, Silver, Bronze), or if they win a medal at all. The following graphic shows the distribution of medals and ages.

Hide

```
AthleteEvents %>%
  ggplot(aes(x = Medal, y = Age)) +
  geom_boxplot(aes(color = Medal)) +
  labs(title = "Distrubtion of Medals by Age") +
  theme_bw()
```



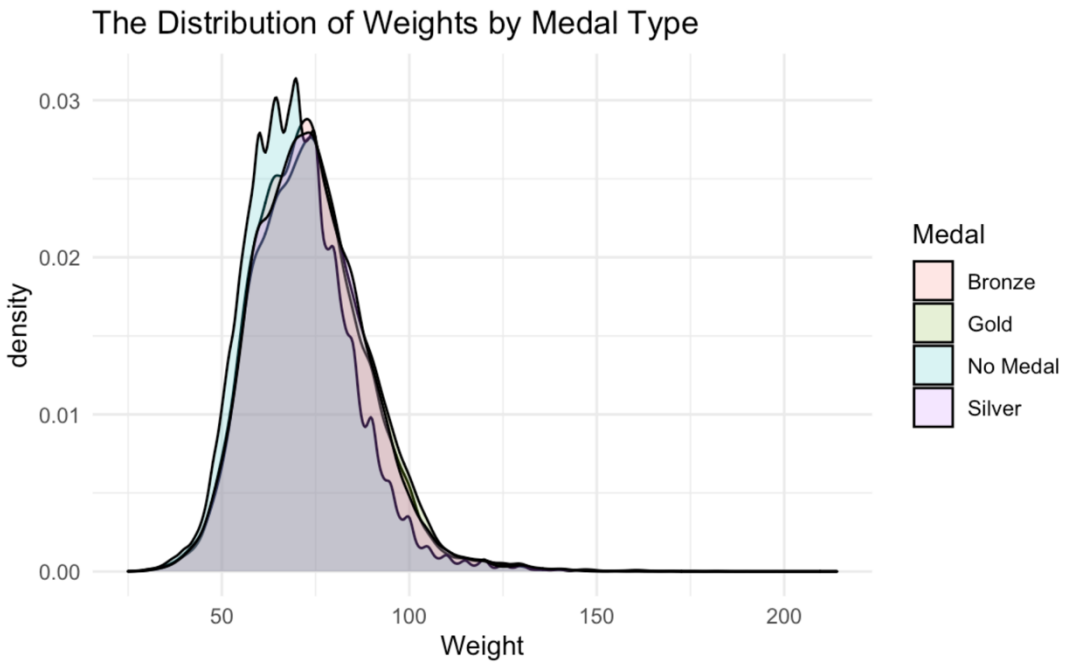
- From the graph above, we can see that the participants of younger ages, particularly in the range of 22-25, win more medals (from Gold, Silver, and Bronze). This was expected because athletes of younger ages are able to win more medals than those athletes who are older. A bit after the age of 37.5, the data contains outliers. Therefore, a boxplot was the best choice for graphing this distribution because it shows the median of the data, as well as the outliers to provide a better analysis.

The Correlation Between Medals and Weight/ Height

- The weight and height of athletes may also influence how many medals they win in the Olympics. It is possible that there is a prime weight and height that is common among athletes who win medals. The following two graphics display the distribution of weight and height among medal types.

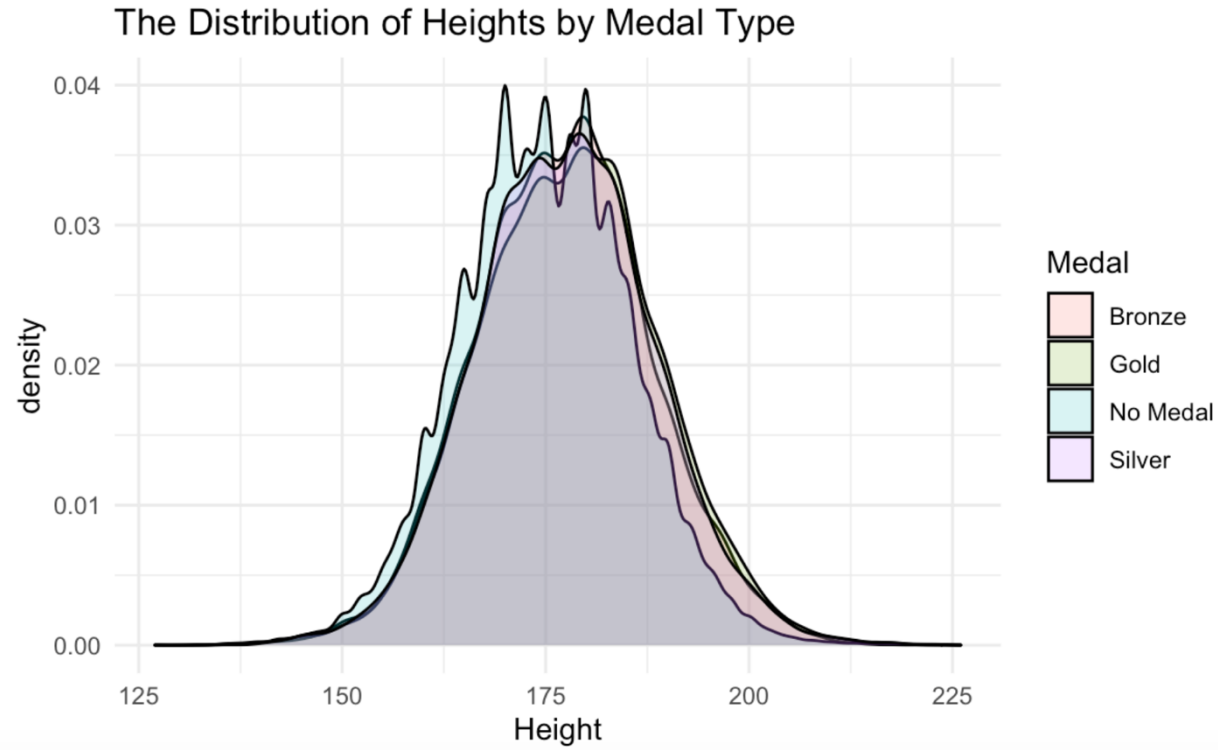
Hide

```
AthleteEvents %>%
  ggplot(aes(x = Weight, fill = Medal)) +
  geom_density(alpha = 0.2) +
  labs(title = "The Distribution of Weights by Medal Type") +
  theme_minimal()
```



Hide

```
AthleteEvents %>%
  ggplot(aes(x = Height, fill = Medal)) +
  geom_density(alpha = 0.2) +
  labs(title = "The Distribution of Heights by Medal Type") +
  theme_minimal()
```



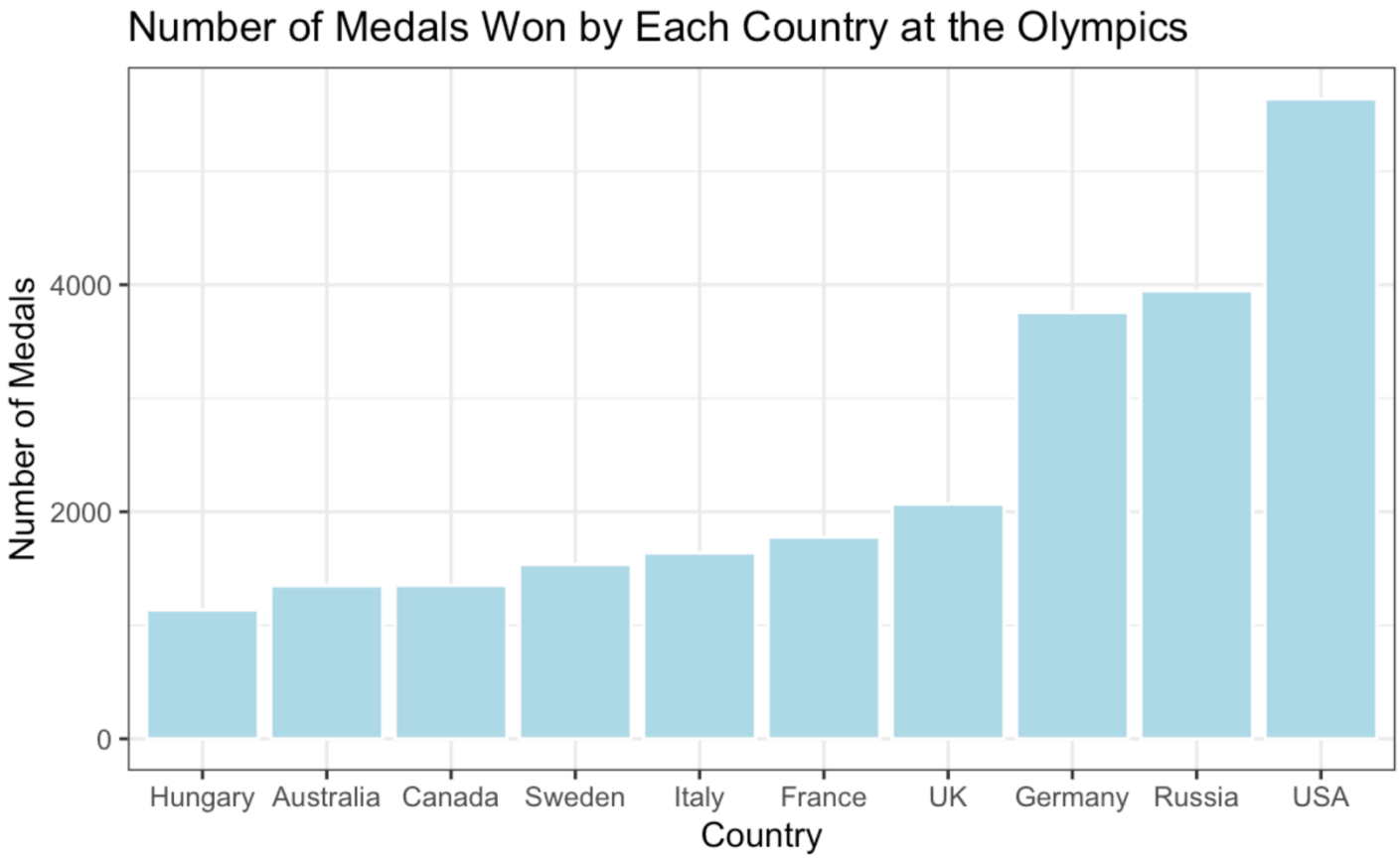
- From the two graphs above, we can see that the most common weight among medal winners is approximately 75 kilograms. The most common height among medal winners is approximately 180 centimeters. While these details about athletes are commonly skipped over when researching different influences of success in the Olympics, they are clearly very important and make an impact on how many medals a country wins.

The Number of Medals Won by Each Country

- Next, we will explore the count of medals won by each country. In the following graphic, we will observe how many medals in total (Gold, Silver, and Bronze) each country has won. The graphic is excluding athletes that did not win a medal at an event. Additionally, only the top 10 countries are shown because a graph with over 200 countries is too difficult to read. Therefore, the top 10 countries with the highest medal count will be observed for the rest of the analysis.

Hide

```
AthleteEvents %>%
  filter(Medal != "No Medal") %>%
  group_by(region) %>%
  summarise(MedalCount = length(Medal)) %>%
  arrange(desc(MedalCount)) %>%
  ungroup() %>% # ungroup() function removes grouping
  mutate(Country = reorder(region, MedalCount)) %>%
  head(10) %>%
  ggplot(aes(x = Country, y = MedalCount)) +
    geom_bar(stat = "identity", color = "white", fill = "light blue") +
    xlab("Country") +
    ylab("Number of Medals") +
    labs(title = "Number of Medals Won by Each Country at the Olympics") +
    theme_bw()
```



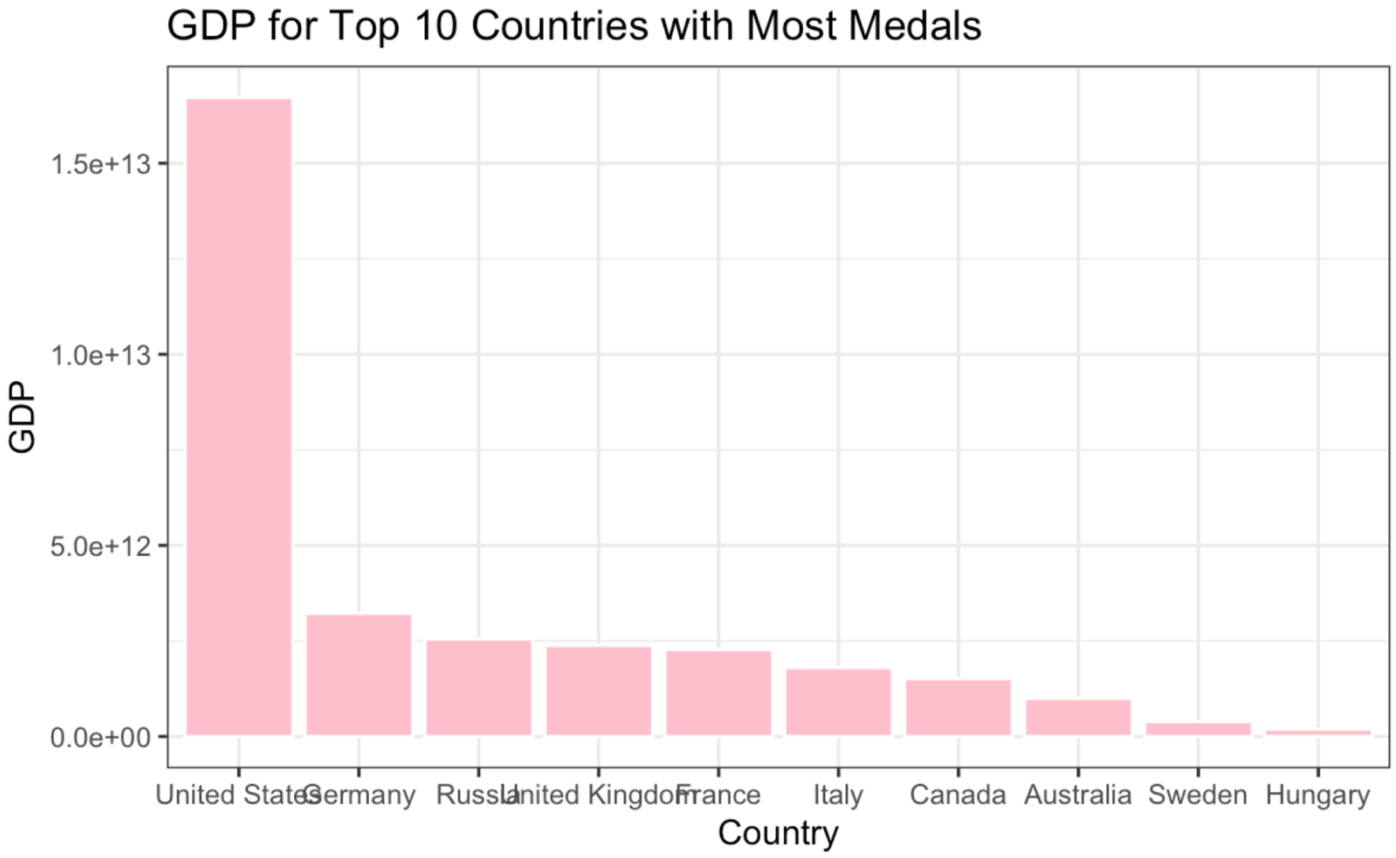
- From the graph above, we can see that the top 10 countries with the most medals over time at the Olympics are: the United States, Russia, Germany, the United Kingdom, France, Italy, Sweden, Canada, Australia, and Hungary. This prompts us to ask the question of why each country has won so many medals, and why did the United States come in first? There are different factors, such as the GDP of a country and the life expectancy/ general health of citizens of a country. These factors can correlate to the athletes' performance in the Olympics. Such factors will therefore be explored in the CountryData data set.

The GDP of the Countries With the Most Medals

- The GDP (Gross Domestic Product) of a country is a valuable economic predictor, and it can be used to evaluate the performance/ amount of medals won by countries in the Olympics. It is likely that countries with a high GDP can afford to invest in training/ athletic programs to help the participants of the Olympics. It is also likely that a high GDP is an indication that citizens have enough money to participate in sports. This may not be the case in countries with less money.
- The CountryData data set contains the GDP for each country in the year 2014. As technology and society progress, it is likely that the GDP for each country increases. The following graphic shows the GDP for the year 2014 of the top 10 countries with the highest medals won in the Olympics.

Hide

```
CountryData %>%
  filter(country == "United States" | country == "Russia" | country == "Germany" | country == "United Kingdom" |
country == "France" | country == "Italy" | country == "Sweden" | country == "Canada" | country == "Australia" | c
ountry == "Hungary") %>%
  arrange(desc(GDP)) %>%
  ggplot(aes(x = reorder(country, desc(GDP)), y = GDP)) +
  geom_bar(stat = "identity", color = "white", fill = "pink") +
  theme_bw() +
  xlab("Country") +
  labs(title = "GDP for Top 10 Countries with Most Medals")
```



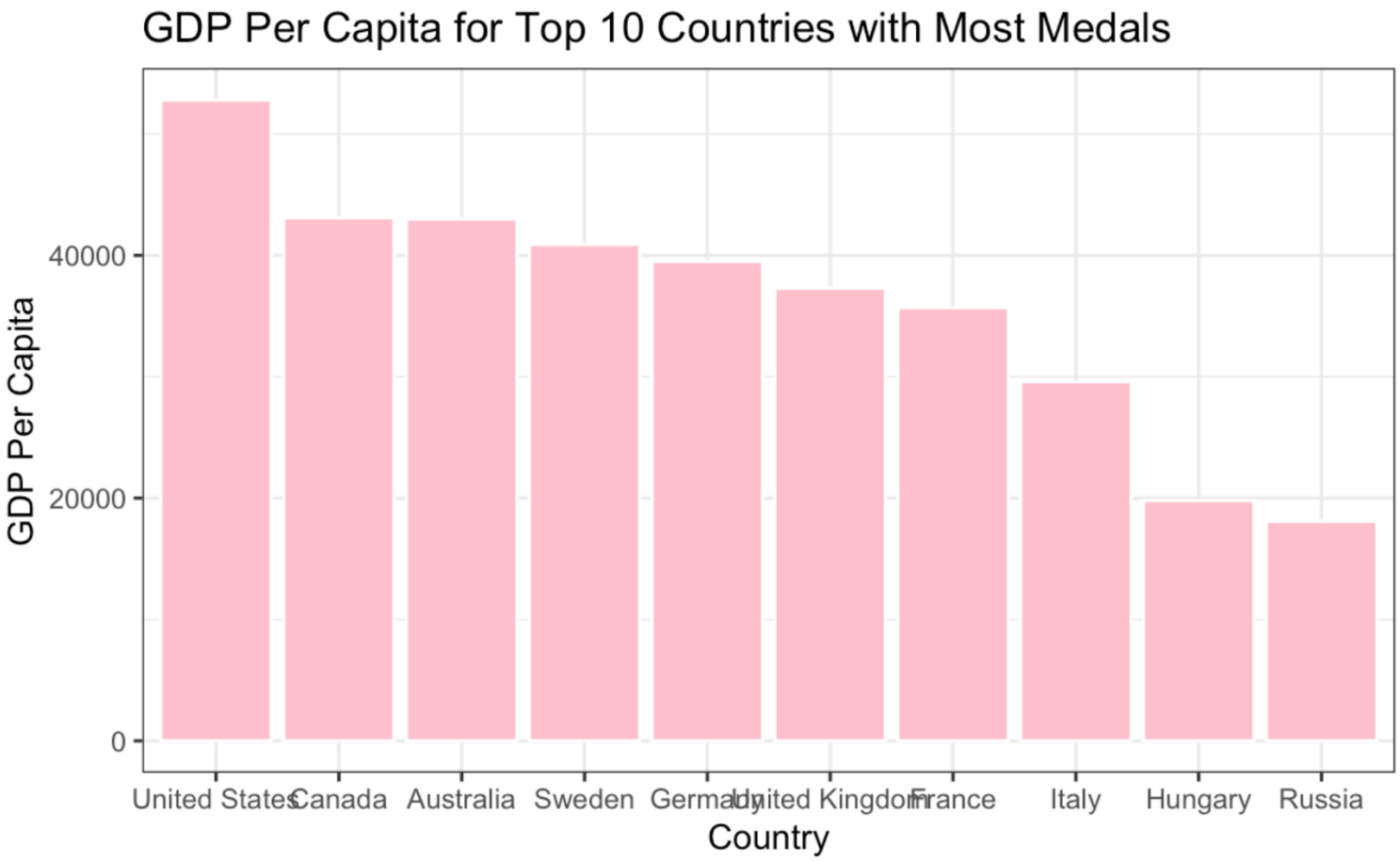
- From the graph above, we can see that the countries with the highest GDPs are, for the most part, in the same order as the countries with the highest medal count for the Olympics. The United States has the highest amount of medals won, and it has the highest GDP. Similarly, Sweden has the lowest medals won among the top 10 countries, and it has the lowest GDP.
- This indicates that there is a correlation among GDP levels and the performance of countries in the Olympics. Wealthier countries have more resources available to train their athletes and guarantee a higher success rate in their events. Alternatively, countries with a lower GDP may not have an abundance of resources, resulting in lower performance rates and medals won in the Olympics.

The GDP Per Capita of the Countries With the Most Medals

- A derivative of the GDP for a country is the GDP per capita. This value gives the earning of an individual person in a country. It is possible that wealthier citizens may have more leisure and resources to participate in sports, so countries with a higher GDP per capita could win more medals in the Olympics. The following graphic shows the GDP per capita for the year 2014 of the top 10 countries with the highest medals won in the Olympics.

Hide

```
CountryData %>%
  filter(country == "United States" | country == "Russia" | country == "Germany" | country == "United Kingdom" |
country == "France" | country == "Italy" | country == "Sweden" | country == "Canada" | country == "Australia" | c
ountry == "Hungary") %>%
  arrange(desc(GDPcapita)) %>%
  ggplot(aes(x = reorder(country, desc(GDPcapita)), y = GDPcapita)) +
  geom_bar(stat = "identity", color = "white", fill = "pink") +
  theme_bw() +
  xlab("Country") +
  ylab("GDP Per Capita") +
  labs(title = "GDP Per Capita for Top 10 Countries with Most Medals")
```



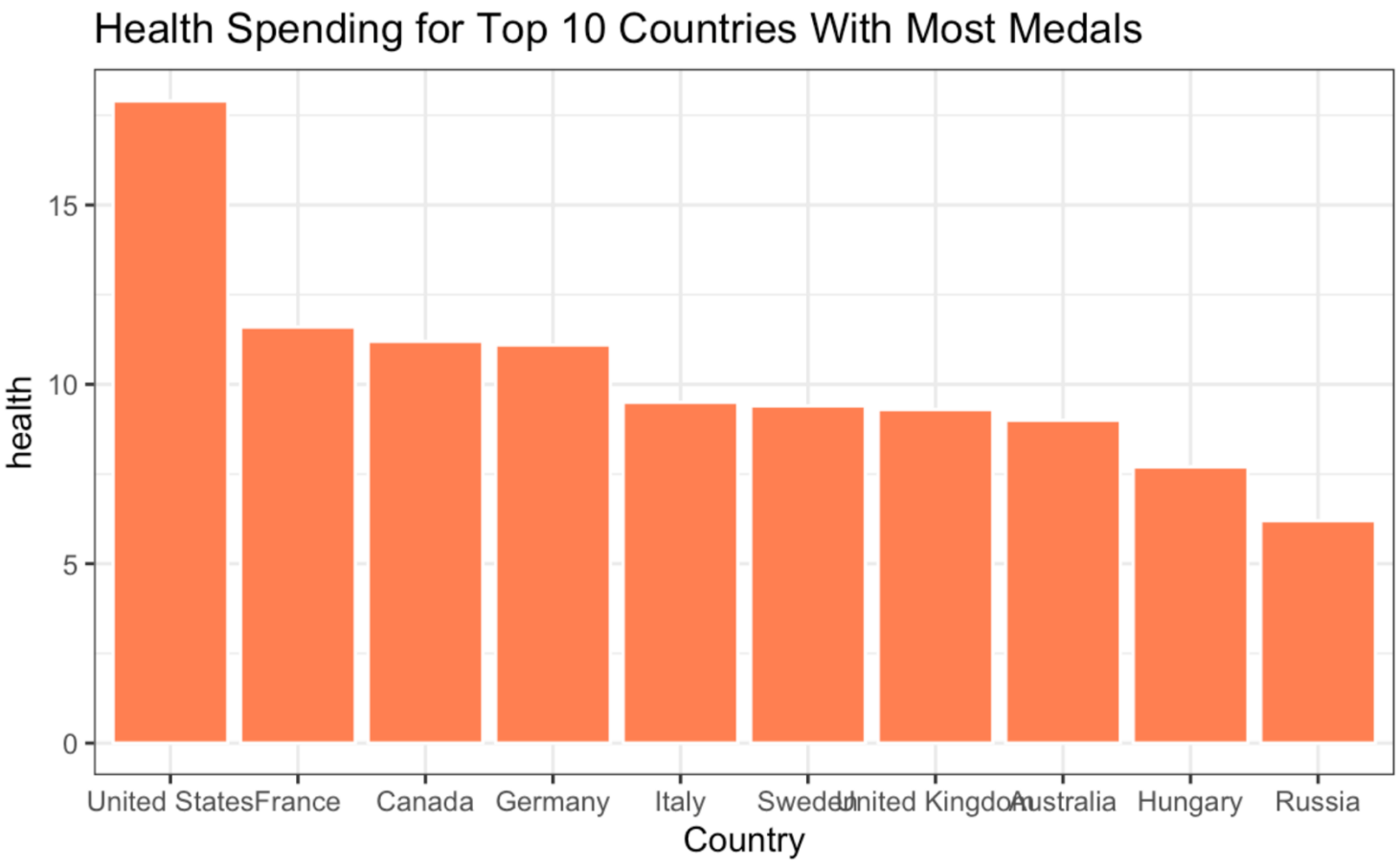
- While the United States once again has the highest GDP per capita along with the most medals won in the Olympics, the graph above shows that there is not much correlation between GDP per capita of a country and the amount of medals they win. The graph shows how countries with lower medal counts have a higher GDP per capita, and countries with high medal counts have a lower GDP per capita. Therefore, we can conclude that the GDP per capita of a country does not influence success rates at the Olympics, but the overall GDP does.

The Health Spending of the Countries With the Most Medals

- Another question I was prompted to ask was whether or not the overall health of participants from certain countries would impact how many medals are won. To explore this, we can study the amount of money that the top 10 countries with the most medals won dedicate to health spending for their citizens. Countries who devote more of their money to the health industry may have healthier citizens, resulting in more efficient/ successful athletes participating in the Olympics. The following graphic displays the health spending of the top 10 countries.

Hide

```
CountryData %>%
  filter(country == "United States" | country == "Russia" | country == "Germany" | country == "United Kingdom" |
country == "France" | country == "Italy" | country == "Sweden" | country == "Canada" | country == "Australia" | c
ountry == "Hungary") %>%
  arrange(desc(health)) %>%
  ggplot(aes(x = reorder(country, desc(health)), y = health)) +
  geom_bar(stat = "identity", color = "white", fill = "coral") +
  theme_bw() +
  xlab("Country") +
  labs(title = "Health Spending for Top 10 Countries With Most Medals")
```



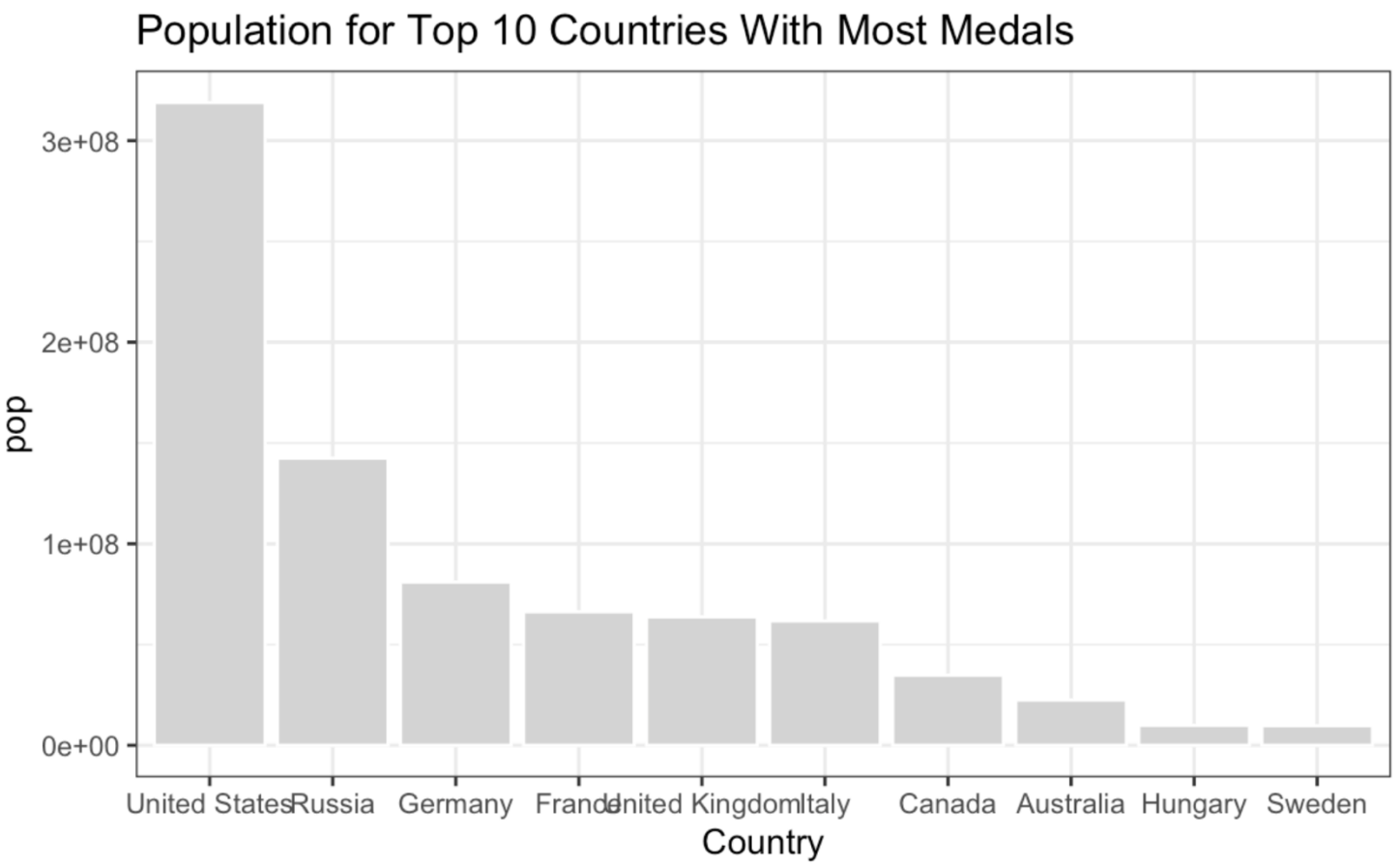
- The United States once again has the highest health spending along with the most medals won in the Olympics. However, the graph shows the countries in a slightly different order compared to the most medals won and highest GDP. Overall, it seems that the amount of health spending for each country relates to the amount of medals they win in the Olympics and their GDP. One surprising detail about the above graph, though, is that Russia has the lowest health spending, but the second highest amount of medals won in the Olympics. Therefore, this outlier may be an indication that the correlation is not as strong as we assumed.

The Population of the Countries With the Most Medals

- The last factor we will observe in the influencing of medals won is the total population of each country. There is a possibility that a larger population of a country would impact how many medals the country wins. This is because a large population leads to a bigger talent pool of athletes to choose from, and choosing the most talented athletes would return higher success rates in the Olympics. The following graphic shows the total population of the top 10 countries with the most medals won.

Hide

```
CountryData %>%
  filter(country == "United States" | country == "Russia" | country == "Germany" | country == "United Kingdom" |
country == "France" | country == "Italy" | country == "Sweden" | country == "Canada" | country == "Australia" | c
ountry == "Hungary") %>%
  arrange(desc(pop)) %>%
  ggplot(aes(x = reorder(country, desc(pop)), y = pop)) +
  geom_bar(stat = "identity", color = "white", fill = "light gray") +
  theme_bw() +
  xlab("Country") +
  labs(title = "Population for Top 10 Countries With Most Medals")
```



- From the graph above, we can see that the countries with the highest populations are, for the most part, in the same order as the countries with the highest medal count for the Olympics. While there are a few countries in a different order, those countries are in their respective ranges in terms of how many medals they have won.
- This correlation indicates that countries with higher populations may have more athletes to choose from, and those athletes who are the most talented are able to represent their country in the Olympics and win a medal (Gold, Silver, or Bronze).

Conclusion

- After observing the top 10 countries that have won the most medals at the Olympics overall, we studied which kinds of factors impacted and correlated with athletes' success rates from these countries. These factors ranged from specific details about the athletes, to economic, health, and social influences from each country. From the analysis of the above graphics, we can answer the research question with the following conclusions:
    - As countries have progressed in various ways, their participation and level of attendance at the Olympics over the years have increased dramatically (in both Winter and Summer). Different influences such as economic distress and the political climate can determine if a country will enter the Olympics. Countries that are more advanced are more likely to enter the events and win more medals at the Olympics.
    - There are three details about athletes that influence how many medals they win: age, weight, and height. For age, we found that the optimal age range for most medal winners (over Gold, Silver, and Bronze) is 22-25 years old. Athletes who do not win any medals are most commonly older. This is a reasonable conclusion because it is expected that younger athletes have higher success rates than their older counterparts. For weight and height, we studied the distribution of these variables by medal type. The graphics showed that most medal winners (across all Gold, Silver, and Bronze medals) weigh 75 kilograms and are 180 centimeters tall. Therefore, there must be a correlation between these physical aspects of athletes and their success in the Olympics.
    - Next, we studied numerous economic factors which can determine how many medals a country wins. We found that the GDP has a much more significant impact than the GDP per capita, especially in the recent years as countries have progressed technologically, politically, and socially. The CountryData data set was an ideal source because it measured the GDP for each country in 2014, which is very recent relative to the AthleteEvents data set. The GDP graph was almost identical to the graph with the top 10 countries with the most medals, indicating that wealthier countries have more resources to train their athletes and succeed in the Olympics. Additionally, we observed that the countries with the highest spending in the health industry correlates to the top 10 countries with the most medals, meaning that the countries who keep their citizens the healthiest produce better turn outs in the Olympic events. Last, we observed that the higher the population of a country, the more medals it wins, most likely due to a higher talent pool of athletes to choose from and represent the country.
    - Overall, we can see that the United States has won the most medals out of any country in the Olympics data set, and consistently proved to have the highest GDP, GDP per capita, health spending, and population. The other countries with the highest medal counts also followed in order for these economic factors. It is very intriguing to learn about what kind of influences can determine the success of countries in the Olympics, and it is necessary to study both the overall status of the athletes, as well as the circumstances each country faces throughout history.