Trends in Olympics from 1896 – 2020

STAT 3355.001

Vinay Polisetty

April 26th, 2023

Table of Contents

1 Introduction 1

2 Data Cleaning 2

Graph 3.1 2

Graph 3.2 2

Graph 3.3 3

Graph 3.4 3

Graph 3.5 3

3 Questions and Findings 3

How Participation Changed in the Last Decade? 4

Which Countries have been the Most Dominating? 5

Number of Athletes by Country 6

Does Age affect Potential to Win? 7

Does GDP have an Affect on Countries Winning 9

4 Conclusions 10

5 Code 11-18

6 References 19

Introduction

The primary purpose of this project was to learn the skills of data analysis by using R programming. For this I was given the task to perform data analysis to discover trends and statistics about the Olympics from 1896 to 2020. The data set provided was retrieved from E-Learning, it had 237,637 observations with 13 variables. These variables included the name of athlete (Name), the gender of the athlete (Sex), the age of the athlete (Age), the team they are representing (Team), the abbreviation form of the team (NOC), the game that they participated in (Games), the year (Year), the Olympic season (Season), the city where the Olympic took place (City), the sport that the athlete played (Sport), the name of the event they participated in (Event), and finally if they won a medal (Medal). To expand on the given dataset, I chose a secondary data set from Kaggle which provides information on the participating countries' GDP from 1960 to 2020. This data set contains 11,507 observations with 4 variables. The four variables were the name of the country (Country), the abbreviation (Country Code), the year for which GDP is reported (Year), and finally the GDP amount (Value). The Olympics have been going on for more than 120 years. Every year more than 200 different countries and 10,000 athletes compete to prove who, and which country is the best. So, there were many different findings to uncover.

In this report I will go over the different findings and thought process in detail. The first step in data analysis is to observe the data given to us and start thinking about the right questions to ask. Since we are working with data over a large time scale it is good to see trends that occurred throughout the years such as changes in participation population. Since there are many countries to observe, we can try to see the progress of different countries over the years. We also have the secondary dataset which helps us to see if there is any correlation between GDP and how that respective country performs in the Olympics.

Data Cleaning

The next most important step to data analysis is data cleaning. Not everything in a data set is useful, there will be times where there are duplicates, null values, or other types of errors. For different graphs we need different kinds of subsets which need to be cleaned so that our process for data analysis is much easier. Normally we can get rid of multiple attributes that might be unnecessary but there might be an issue where we might have to go back to the original dataset hence, I wanted to make different subsets for each graph and keep the original/raw data set as it is in case if I must go back to it.

Graph 3.1

For the first graph, Fig 3.1 Population for The Last Decade, the first step was to minimize the dataset only for the years 2000 – 2020. The reason for only the past few years is because of the huge number of years that the Olympics has been going on for, hence the last decade sounds like a reasonable number that captures both the recent/past trends in the Olympics. Once we subset it to only the past decade we need to get rid of unnecessary columns and only keep gender and year. I created two subsets for the male and female data between 2000 and 2020 and got rid of any null values. This can be done by running *unique.(col_name)* and we see 'NA' we know there is a null value and if not, our data set is clean.

Graph 3.2

For the next graph, Fig 3.2 "Top 10 Country winnings", this graph is to show the countries that won the greatest number of medals throughout the years. To get the subset for this we first need to take only the rows which won medals. So, we exclude all the null values in the medal column. Then we need to find the top 10 countries. We do this by creating a loop to count the number of medals each country won and take those countries and make another subset. Finally, we can use this to create a horizontal bar plot.

Graph 3.3

This graph didn't need much cleaning. This graph is to basically show where most of the athletes come from. I made a subset to count the number of athletes from each country and then combining it with a world dataset to plot in on a map. I got rid of unnecessary columns such as medal, NOC, age, and name of athlete. The issue that arose was that the column names for the world dataset and the subset didn't match. Once this problem was solved, we can perform merge operation on both the subsets to get our required subset for mapping.

Graph 3.4

This graph shows how age might be related to the number of medals an athlete won. First, we needed to find how many different age groups there were. So, we first subset the data by only keeping the age column and medal column without null values. There were found to be null values in the age column, so we use complete.cases() to get rid of 'NA'. Once we find the unique age groups, we save it in a vector and then later use it to count the number of medals they won. This subset is then plotted on a scatter plot.

Graph 3.5

This final graph is to see if there is a correlation in the GDP of a country to the number of medals they've won. Luckily our GDP dataset didn't have any null values so there was no cleaning required. As for our Olympics dataset I made a subset of only the countries that won medals. Once that is done, we take only the top 10 countries that won the highest number of medals. We later merge this with the GDP dataset. Once again, we have the issue of column names not matching and United States being USA. Once we have them to match, we can just merge our two datasets to use for plotting. Since the GDP dataset starts from 1960 instead of 1896, I decided to only use the GDP for the last decade.

Questions and Findings

Once we have our questions and datasets ready to go we need to know plot/visualize it so that we can answer the questions we have formed.

How Participation Population Changed in The Last Decade?

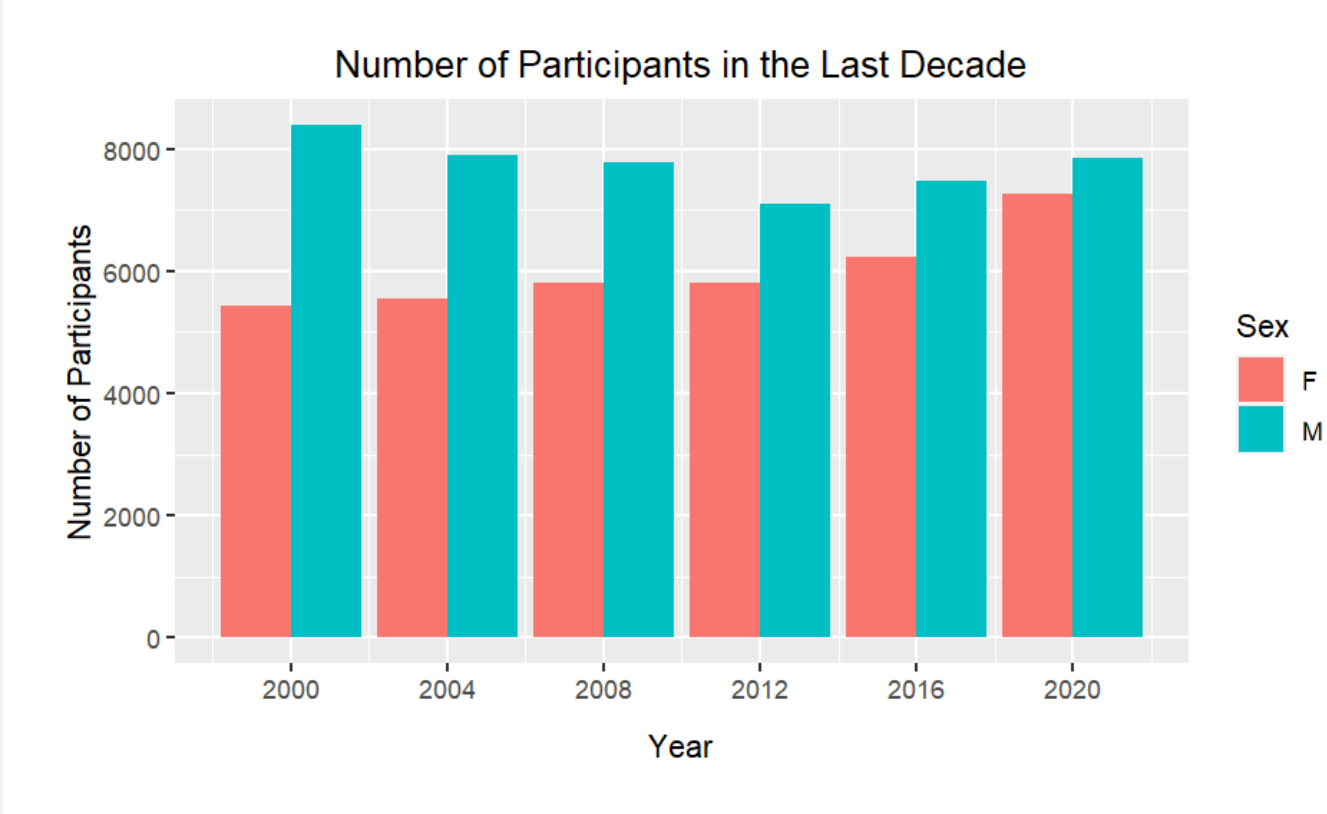

Figure 3.1 Number of Participants in the Last Decade

During the first Omlympic event the number of participants was only a mere 241athletes. This number has been exponentially increasing and in the most recent 2020 Tokyo Olympics there were 11,500 athletes participating. One interesting thing to see is how the participant population is divided in terms of gender. Until the 2000's the percentage of female athletes to male was very low 30%. But from the 2000's it has changed significantly. In 2000 they made up 38% of the total partcipants. And from the graph we can see how that continues to increase. We can also observe that how there was a small decline in males, but it goes back up again. During this time the female participation rose to being 48.7% of the participants. This is the most balanced the Olympics has ever seen. There might be many causes for, but it can mainly be explained by the

increased number events for females. Many countries have also increased its funding into female athletes which might encourages more females to join the Olympics.

Our next graph is to answer the question of which conutry has been the most dominating over the years.

Which Countries Have Been the Most Dominating

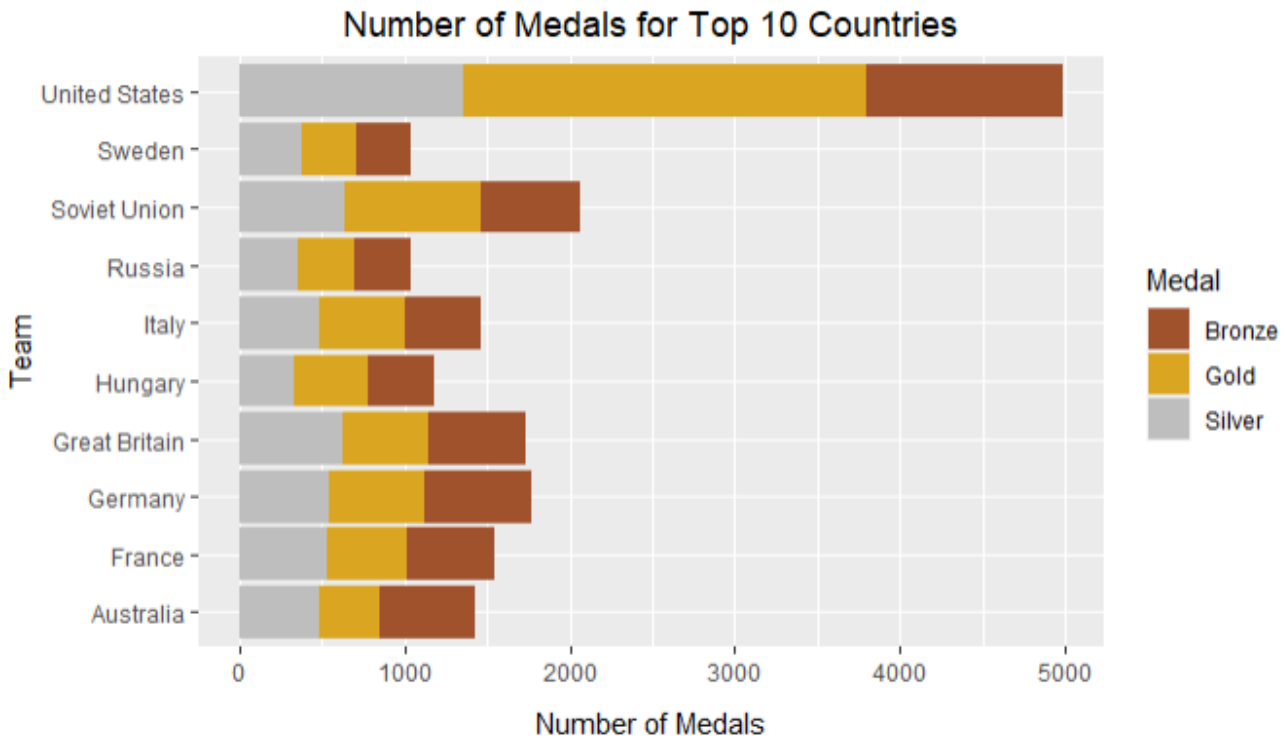

Figure 3.2 Number of Medals for Top 10 Countries

In some form the winners of the Olympics represent how strong their country is. It was interesting to see which countries have been the most dominating over the past 100 years. Throughout history United States of America and Russia/Soviet Union have been known as world superpowers because of the strength in their economy and military. This pattern can also be seen in the Olympics where the US has been the most dominating by winning almost 5,000 medals. And as for Soviet Union/Russia it is a total of around 3000 medals. The list continues as Great Britain, Germany, France, Australia, Sweden, and Hungary. It might seem that it is many medals, but this is because it is the medal for the number of participants than number of teams.

Because when they count the number of medals a country won they count by teams. For example, if the US basketball team won a gold medal, it counts as one. But since there are five players in a team, this dataset contains 5 gold medals instead of just one because it went by the athlete's name. Even with this in consideration the list remains the same where the US is the dominating country when it comes to winning medals in the Olympics. We can also take this as a snowball effect because once you start winning many medals it is empowering to the team and the country as whole, which will encourage the country to spend more on their team to improve their results.

Number of Athletes by Country

The next graph shows which countries the majority of the atheltes come from on a world map.

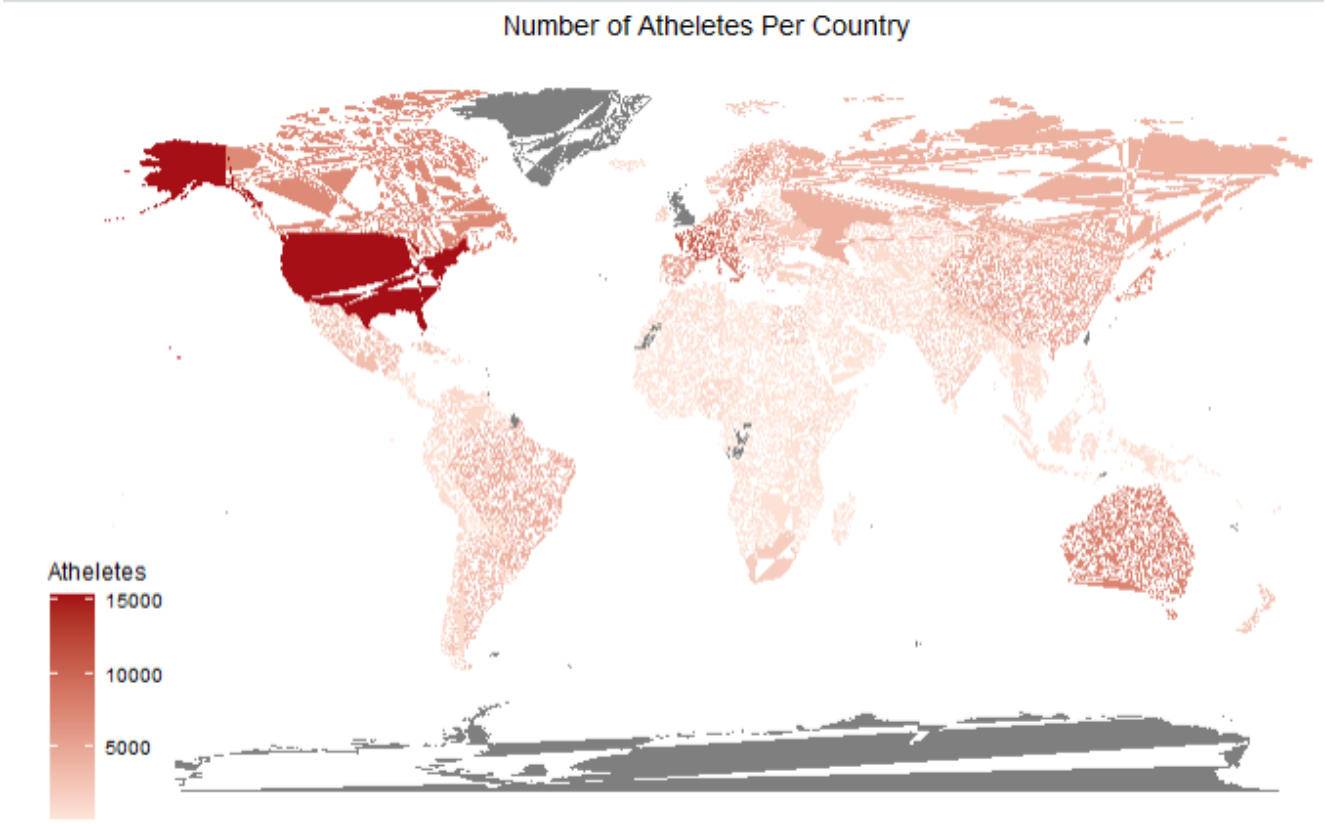

Figure 3.3 Number of Athletes by Country

Instead of a normal bargraph this is an interesting way to represent which country the athletes come from. It makes it easier to understand, due to the gradient fill where darker the red the greater number of athletes and lighter the red a smaller number of atheltes. With that it is clearly

distinguishable that the United States had themost number of atheltes throughout the years with almost 15,000 participants. Next comes the european countries such as the Great Britain, France, Germany with around 10,000 participants each. China and Russia follow next with aroun 8,000 atheltes as well as Australia. The remaining countries make up a very few portion alsmost less than 5,000 athletes. One might think that wouldn't more population equate to more athletes, but unfortunately is not the case. Population is just one factor, but the biggest factor is money. Countries with more money have the potential to train more athletes. They have the infrastructure and transportation to send athletes all over the world. We will be seeing how money might affect winnings later in the report.

Does Age Affect Potential to Win

As time goes we see that our favorite athlete doesn't play the same as he used to. He isnt as explosive or strong anymore. They don't win as much because they run out of stamina quickly and to sum it up they are getting old. At the same time, we can argue that young athletes don't perform as well because they lack experience. So, we know that age clearly has an affect we also want to know what age groups a clear difference is there. Hence I made a scatter plot to visualize where most of the medals are won.

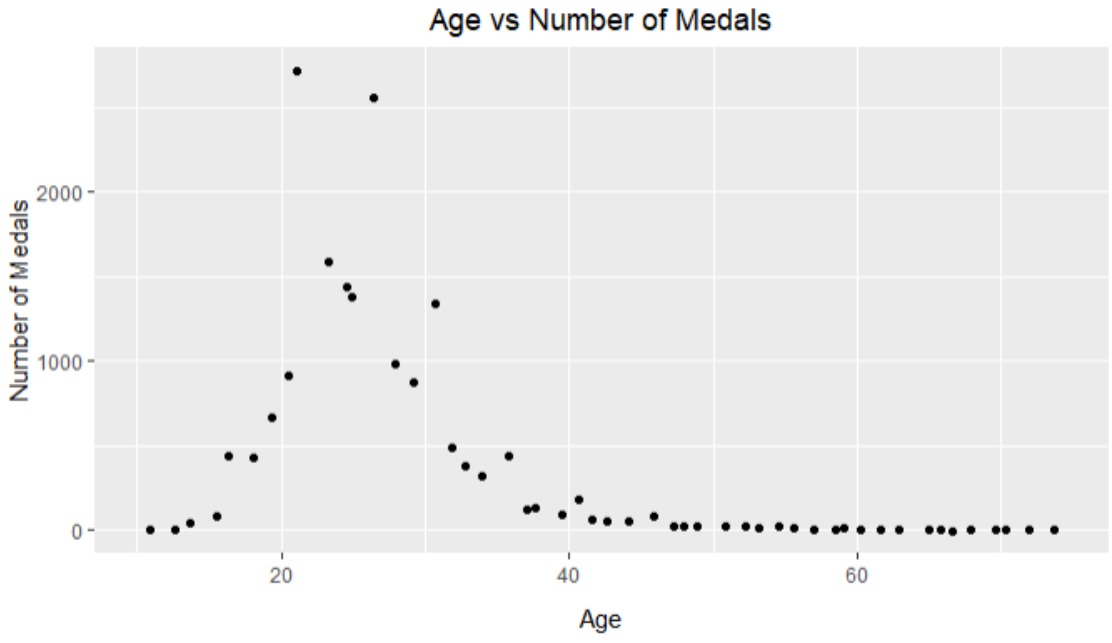

Figure 3.4 Age vs Number Medals

From figure 3.4 we observe that we get a bell curve type graph or also called a normal distribution graph. This is where the ends of the curve are lower compared to the middle. In our case this meas that the younger athletes from ages 11 – 16 have barely won any Olympic medals. This is because of the lack of experience as well as the number of participants in that age group. And we can say the same for ages 35 and above. Where even though they have a good amount of experience they are bound by physical limitations due to their age. Finally, when we observe the middle, which is age group 19 – 32, this is where most of the medals are won. This is because they are in their "prime". This is where they have their peak athletic performance as well as a good amount of experience. Added to this most of the athletes are from this age group hence we see more medals. These kinds of models might help countries to decide where their funding goes. This means they can try to provide more funding to middle age groups where they've seen the most success. They also try to provide funding for ages between 17-18 so that when they reach their "prime" age they might be able to outperform older generations.

Our last graph is going to show if there is any relation between how rich a country is to the number of medals they have won.

Does GDP have an Affect on Countries Winning

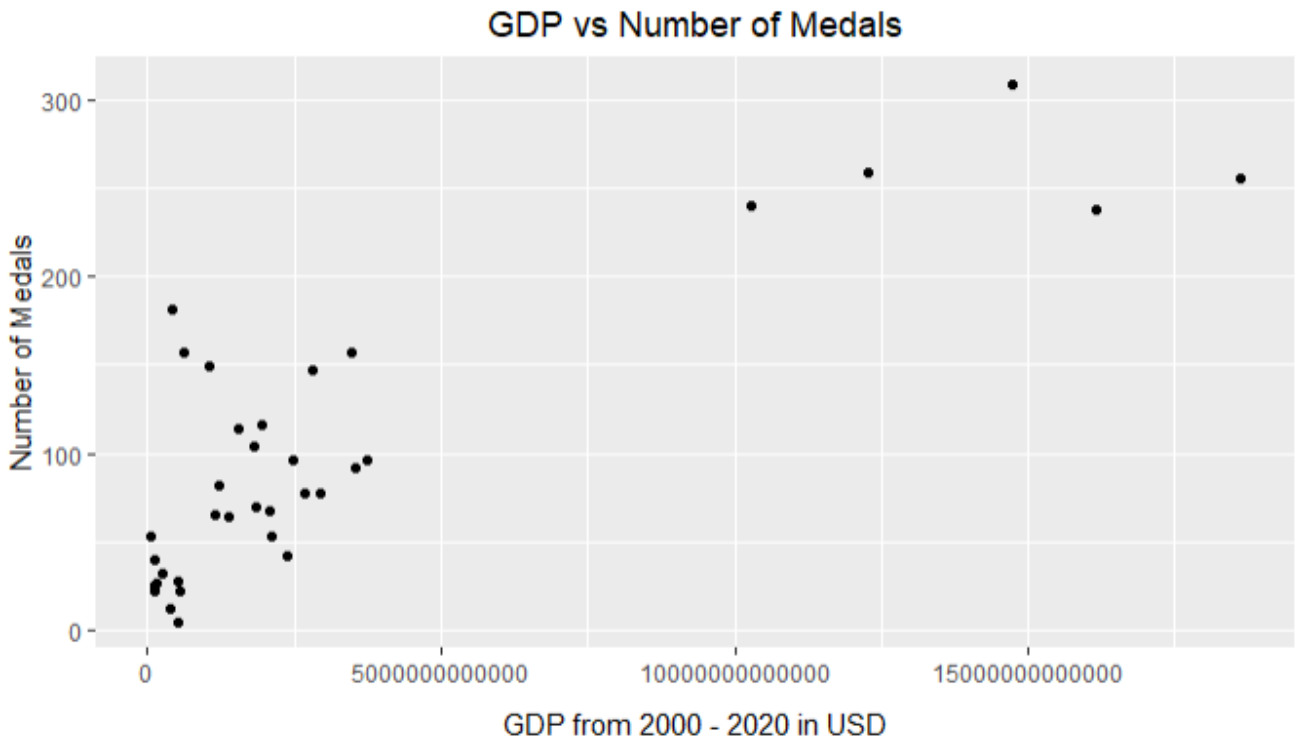

Figure 3.5 GDP vs Number of Medals

GDP means gross domestic product. GDP is the measure of how the economic growth of a country is which basically means higer GDP means richer country. Having a higher GDP means the country can spend more supporting their Olympic team. First I chose to take GDP for the past decade as too many years would've made the graph messy. But this graph depicts the trend that is accurate. We can see that countries with a higher GDP tend to outperform the ones with lower GDP. Since this is a bit too broad I narrowed it down to only the top ten countries that won medals. Even then the graph follows the pattern above where the higher the GDP the more likely they are to win medals. We can also observe there are very few countries with a GDP of over 10 trillion dollars. These are countries like the US, Great Britain, and Russia. This is also the reason why we see the majority of the athletes coming from these countries from Figure 3.3. We have some outliers in the lower GDP countries, this means their athletes outperformed and beat expectations.

Conclusions

Through this project we were able to go through the steps of proper data analysis by first asking the right questions. Once this is done we clean the data set and add any secondary data set that might enhance our understanding. Finally, we can answer these questions in this report and make visualizations to make it easier to understand.

In this project I worked with the Olympics data set and added a secondary GDP data set to form informative questions. The main questions that I asked were how the participation population changed over the years. We were able to observe that the overall population has been in an upward trend. And taking gender into account we see that female population has been rising gradually which is a good sign. It would also be interesting to see which countries have been the most dominating hence the second question. And from the graphs we see that the USA outperforms all countries by a huge margin, almost having 4,000 medals.

The third question was to see from which country the majority of the athletes come from. And once again we see that the USA has the greatest number of athletes, around 15,000 athletes throughout the years. This also explains why they have been winning the majority of medals. USA is followed by countries like the UK, Germany, Great Britain, China, and Russia. These countries also fall under the countries that won the greatest number of medals. Hence we can conclude that there is some relation between the number of athletes and the number of medals a country wins. We were able to visualize this in the form of a wold map.

Questions 4 and 5 try to find a relation if age has anything to do with number of medals. Question 5 tries to find relation between GDP of a country and number of medals won. In question 4 we conclude that age does have a factor and we see the majority of the athletes winning between the age group of 19 – 32. And from question 5 we conlcude that the higher the GDP of a country the more chances they have at winning an Olympic medal.

Hence I was able to succesfully show the process of data analysis on the Olympics and GDP data set.

Code

```
#Installing Required libraries

library(ggplot2)

library(maps)

library(ggthemes)


#Ingesting Raw data

raw_data <- read.csv("olympics.csv")


#Dividing dataset by gender

male_data <- raw_data[raw_data$Sex == "M",c("Name", "Sex", "Age", "Team",  "Year", "City", "Sport", "Event", "Medal")]

female_data <- raw_data[raw_data$Sex == "F",c("Name", "Sex", "Age", "Team",  "Year", "City", "Sport", "Event", "Medal")]


#Creating data set for last decade

lastdec_data <- raw_data[raw_data$Year %in% c(2000, 2004, 2008, 2012, 2016, 2020), c("Year","Sex") ]


#Creating plot for Figure 3.1 (Population change over the years)

ggplot(data = lastdec_data) + geom_bar(mapping = aes(x = Year, fill = Sex), position = "dodge")

+ labs(y = "Number of Participants")

+ theme(plot.margin = unit(c(1, 1, 2, 1), "lines"),axis.title.x = element_text(margin = margin(t = 10)), plot.title = element_text(hjust = 0.5))

+ ggtitle("Number of Participants in the Last Decade")

+ scale_x_continuous(breaks = c(2000, 2004, 2008, 2012, 2016, 2020))


#Creating subset for medals data

medal_data <- raw_data[raw_data$Medal %in% c("Gold", "Silver", "Bronze"), c("Team", "Medal", "Age", "Year")]
```

```
#Finding all unique countries
all_countries <- c(unique(raw_data$Team))
countries <- c(unique(medal_data$Team))
count_part <- numeric(1169)
count_medal <- numeric(486)

#Loop for counting number of medals each country won
for(i in 1:486){
  count_medal[i] <- nrow(medal_data[medal_data$Team == countries[i],])
}

#Data frame with country and respective medal counts
country_medal_df <- data.frame(countries,count_medal)

#Sorting Data to find top 10 countries
data_sorted <- country_medal_df[order(-country_medal_df$count_medal),]
top_10 <- head(data_sorted, n = 10)
top_10_medals <- medal_data[medal_data$Team %in% c(top_10$countries),]

#Creating Figure 3.2(Top 10 countries winnings)
ggplot(data = top_10_medals) + geom_bar(mapping = aes(y = Team, fill = Medal))
+ scale_fill_manual(values = c("Gold" = "goldenrod", "Silver" = "gray", "Bronze" = "sienna"))
+ labs(x = "Number of Medals", y = "Team")+ theme(plot.margin = unit(c(1, 1, 2, 1),
"lines"),axis.title.x = element_text(margin = margin(t = 10)), plot.title = element_text(hjust =
0.5))
+ ggtitle("Number of Medals for Top 10 Countries")

#Preparing clean data without null values from medal data
```

```
clean_medal <- medal_data[complete.cases(medal_data$Age), ]

clean_medal <- clean_medal[clean_medal$Team %in% c(top_10$countries),]

age_groups <- unique(na.omit(medal_data$Age))


#Creating plot for figure 3.4(Does age matter to win medals)

ggplot(clean_medal, aes(x = Age, y = ..count..)) + geom_point(stat = "bin", bins = 50, position =
"jitter")

+ labs(x = "Age", y = "Number of Medals")+ theme(plot.margin = unit(c(1, 1, 2, 1),
"lines"),axis.title.x = element_text(margin = margin(t = 10)), plot.title = element_text(hjust =
0.5))

+ ggtitle("Age vs Number of Medals")


#loop for counting number of athletes from each country

for(n in 1:1169){

  count_part[n] <- nrow(raw_data[raw_data$Team == all_countries[n],])

}


count_athelete <- data.frame(all_countries,count_part)

count_athelete$all_countries <- gsub("United States", "USA", count_athelete$all_countries)


#Reading world map data

world <- map_data("world")

df_map <- merge(world, count_athelete, by.x = "region", by.y = "all_countries", all.x = TRUE)


color_scale <- scale_fill_gradient(low = "#FEE5D9", high = "#A50F15")


#Creating World Map for figure 3.3

ggplot(data = df_map) + geom_polygon(mapping = aes(long, lat, group = group, fill =
count_part))
```

```
+ scale_fill_gradient(low = "#FEE5D9", high = "#A50F15", guide = "legend") + labs(fill = "Atheletes") + theme_map()

+ color_scale + theme(plot.title = element_text(hjust = 0.5)) + ggtitle("Number of Atheletes Per Country")


#Ingesting Raw GDP data set

gdp_raw <- read.csv("gdp.csv")


#Creating GDP dataset for last decade

gdp_last_dec <- gdp_raw[gdp_raw$Year %in% c(2000,2004,2008,2012,2016,2020), c("Country.Name", "Year", "Value")]

gdp_last_dec_top_10 <- gdp_last_dec[gdp_last_dec$Country.Name %in% c(top_10$countries),]

colnames(gdp_last_dec_top_10) <- c("Country", "Year", "Value")


#Merging the GDP dataset with medal data set

gdp_medal_merge <- merge(country_medal_df, gdp_last_dec_top_10, by.x ="countries", by.y = "Country")


medal_data_aus <- top_10_medals[top_10_medals$Team == "Australia",c("Team","Medal","Year")]

medal_data_aus_dec <- medal_data_aus[medal_data_aus$Year %in% c(2000,2004,2008,2012,2016,2020),]


count_aus <- numeric(6)

dec_years <- c(2000,2004,2008,2012,2016,2020)

for(a in 1:6){

  count_aus[a] <- nrow(medal_data_aus_dec[medal_data_aus_dec$Year == dec_years[a],])

}


#Counting number of medals each country won
```

```
medal_data_fran <- top_10_medals[top_10_medals$Team ==
"France",c("Team","Medal","Year")]

medal_data_fran_dec <- medal_data_fran[medal_data_fran$Year %in%
c(2000,2004,2008,2012,2016,2020),]


count_fran <- numeric(6)

for(b in 1:6){

  count_fran[b] <- nrow(medal_data_fran_dec[medal_data_fran_dec$Year == dec_years[b],])

}


medal_data_ger <- top_10_medals[top_10_medals$Team ==
"Germany",c("Team","Medal","Year")]

medal_data_ger_dec <- medal_data_ger[medal_data_ger$Year %in%
c(2000,2004,2008,2012,2016,2020),]


count_ger <- numeric(6)

for(c in 1:6){

  count_ger[c] <- nrow(medal_data_ger_dec[medal_data_ger_dec$Year == dec_years[c],])

}


medal_data_hun <- top_10_medals[top_10_medals$Team ==
"Hungary",c("Team","Medal","Year")]

medal_data_hun_dec <- medal_data_hun[medal_data_hun$Year %in%
c(2000,2004,2008,2012,2016,2020),]


count_hun <- numeric(6)

for(d in 1:6){

  count_hun[d] <- nrow(medal_data_hun_dec[medal_data_hun_dec$Year == dec_years[d],])

}
```

```
medal_data_ita <- top_10_medals[top_10_medals$Team == "Italy",c("Team","Medal","Year")]

medal_data_ita_dec <- medal_data_ita[medal_data_ita$Year %in%
c(2000,2004,2008,2012,2016,2020),]


count_ita <- numeric(6)

for(e in 1:6){

  count_ita[e] <- nrow(medal_data_ita_dec[medal_data_ita_dec$Year == dec_years[e],])

}


medal_data_swe <- top_10_medals[top_10_medals$Team ==
"Sweden",c("Team","Medal","Year")]

medal_data_swe_dec <- medal_data_swe[medal_data_swe$Year %in%
c(2000,2004,2008,2012,2016,2020),]


count_swe <- numeric(6)

for(f in 1:6){

  count_swe[f] <- nrow(medal_data_swe_dec[medal_data_swe_dec$Year == dec_years[f],])

}


medal_data_usa <- top_10_medals[top_10_medals$Team == "United
States",c("Team","Medal","Year")]

medal_data_usa_dec <- medal_data_usa[medal_data_usa$Year %in%
c(2000,2004,2008,2012,2016,2020),]


count_usa <- numeric(6)

for(g in 1:6){

  count_usa[g] <- nrow(medal_data_usa_dec[medal_data_usa_dec$Year == dec_years[g],])

}


country_aus <- rep("Australia", times = 6)
```

```
country_fran <- rep("France", times = 6)
country_ger <- rep("Germany", times = 6)
country_hun <- rep("Hungary", times = 6)
country_ita <- rep("Italy", times = 6)
country_swe <- rep("Sweden", times = 6)
country_usa <- rep("United States", times = 6)

#Creating table columns for each country
count_year_aus <- data.frame(dec_years, rep("Australia", times = 6),count_aus)
colnames(count_year_aus) <- c("Year", "Country", "count")
count_year_fran <- data.frame(dec_years,rep("France", times = 6),count_fran)
colnames(count_year_fran) <- c("Year", "Country", "count")

count_year_ger <- data.frame(dec_years,rep("Germany", times = 6),count_ger)
colnames(count_year_ger) <- c("Year", "Country", "count")

count_year_hun <- data.frame(dec_years,rep("Hungary", times = 6),count_hun)
colnames(count_year_hun) <- c("Year", "Country", "count")

count_year_ita <- data.frame(dec_years,rep("Italy", times = 6),count_ita)
colnames(count_year_ita) <- c("Year", "Country", "count")

count_year_swe <- data.frame(dec_years,rep("Sweden", times = 6),count_swe)
colnames(count_year_swe) <- c("Year", "Country", "count")

count_year_usa <- data.frame(dec_years,rep("United States", times = 6),count_usa)
colnames(count_year_usa) <- c("Year", "Country", "count")
```

```
#Combining merged data set with all table columns

combined_df <- rbind(count_year_aus, count_year_fran, count_year_ger, count_year_hun,
count_year_ita, count_year_swe, count_year_usa)


#Creating GDP vs Number of medals plot from figure 3.5

gdp_plot <- merge(combined_df,gdp_last_dec_top_10, by.x = c("Year", "Country"), by.y =
c("Year", "Country"))

options(scipen = 999)

ggplot(gdp_plot, aes(x = Value, y = count)) + geom_point()

+ labs(x = "GDP from 2000 - 2020 in USD", y = "Number of Medals")

+ theme(plot.margin = unit(c(1, 1, 2, 1), "lines"),axis.title.x = element_text(margin = margin(t =
10)), plot.title = element_text(hjust = 0.5))

+ ggtitle("GDP vs Number of Medals")
```

References

"Women Participants in Olympic Summer Games from 1900 to 2020." Statista, Statista Inc., 2021, https://www.statista.com/statistics/531146/women-participants-in-olympic-summer-games/. Accessed 26 Apr. 2023.

Minsberg, Talya. "As More Olympic Athletes Come Out, Team Officials Navigate New Territory." The New York Times, 22 July 2021, https://www.nytimes.com/2021/07/22/sports/olympics/olympics-athletes-gender.html.