Summer Olympic Medals

Jonathan Lynch

data: https://www.kaggle.com/divyansh22/summer-olympics-medals

The Summer Olympic Medals dataset evaluated within this report is a fairly large dataset consisting of over 15,300 observations, and 14 variables/features (most of which are categorical in nature). Each observation represents a single medal won at the Summer Olympic Games from 1976 to 2008. Some of the variables included are the year, sport, discipline, event, athlete name, gender, medal type (gold, silver, or bronze), and the country represented. A variety of visualization techniques are explored within this report in order to appropriately display different aspects of the data. Some of these techniques include a treemap, a stacked bar chart, a choropleth, a small multiples dodged bar graph, as well as an animated plot. Each technique was chosen to help illustrate a specific aspect of the data, as well as convey a compelling story as a whole relating to this particular dataset.

Initially, a handful of exploratory visualizations were created in order to obtain a more complete understanding of the relationships between some of the variables included. For instance, the variable gender was explored in several ways. First, the participation rates for both male and female athletes were compared over time using a line graph. Next, the aggregate number of medals won over all Summer Olympic years for both male and female competitors was broken down by individual sport through a stacked bar chart. A stacked bar graph was also created for the purpose of exploring which countries won the most medals specifically in the sport of boxing.

Eventually, the story begins by analyzing the breakdown of sports by discipline and number of medalists through the use of a treemap. This effectively illustrates which Summer Olympic sports were the most popular in terms of the overall number of Olympic medalists per sport. Next, the number of individual medals won (gold, silver, and bronze) for each of the top ten athletes is evaluated through a stacked bar chart. A choropleth displaying the total number of medals won for all countries is then interpreted to gain a solid understanding of which nations have historically been the most successful at the Summer Olympics. Finally, the counts of each specific medal type (gold, silver, and bronze) are compared over time for the top two countries via a small multiples dodged bar graph. Additionally, a link to an animated plot illustrating the total number of medals won over time for the top six countries is included at the end in order to provide a high level comparison of a handful of the top performing nations.

Figure 1: Treemap of Largest Summer Olympic Sports

The treemap in Figure 1 illustrates the breakdown of sports by discipline and number of medal winning athletes for all sports held at the Summer Olympic Games from 1976 to 2008. The larger and darker each square/rectangle on this treemap, the greater the number of Summer Olympic medalists that participated in that particular sport/discipline. This graph is informative because it gives a general idea about the popularity of each sport in terms of the total number of Olympic athletes that received medals in it over the course of nine Summer Olympics. From this plot, it is apparent that the largest/most popular sports were Athletics, Aquatics -- swimming, and Rowing (approximately 1,500 medalists each). Some of the less popular sports were Tennis, Badminton, and Modern Pentathlon (less than 500 medalists).

In this visualization, both color and area were mapped to the number of medalists. A continuous color scale from light blue to dark blue was utilized, and the text color of red was selected due to the fact that it is located on the opposite side of the color wheel as blue (and thus does not easily blend together). The white text also stands out, but was made slightly transparent through the adjustment of its opacity level in R. This visualization was refined through the drafting process by reversing the transition of the continuous color scale, which initially defaulted to scale from dark (low value) to light (high value), instead of light to dark.

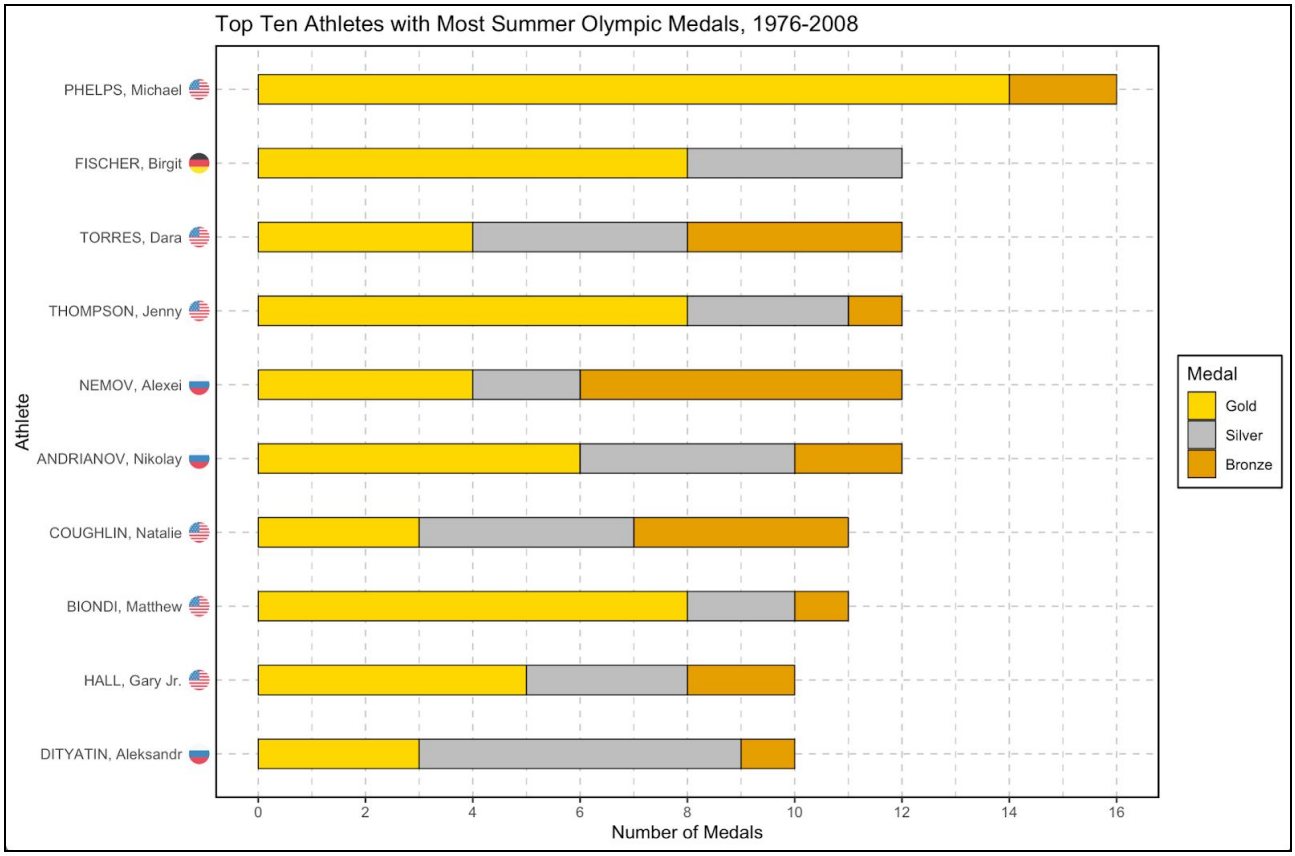

Figure 2: Stacked Bar Graph of Top Ten Summer Olympic Athletes

The stacked bar graph in Figure 2 shows the number of medals won for the top ten athletes at the Summer Olympics from 1976 to 2008, broken down by specific medal type (gold, silver, and bronze). One theme that begins to emerge with this visualization is that of the top ten athletes who competed in the Summer Olympics over approximately three decades, the majority were represented by just two countries, the United States and Russia. From this graph, we can also see that Michael Phelps, the famous American swimmer, won the most Summer Olympic medals (14 gold, and 2 bronze) between 1976 and 2008. The treemap in Figure 1 illustrated that Aquatics -- swimming was one of the most popular Summer Olympic sports/disciplines. Aquatics also happens to be the sport with the greatest number of events. Thus, it intuitively makes a great deal of sense that the top medal winning athlete might be a swimmer.

In this visualization, color was appropriately mapped to the specific type of medal won (gold, silver, or bronze). The plot axes were flipped in order to allow for the names of the individual athletes to be more easily read, and through the drafting process the ggflags library in R was employed to create the round flag glyphs illustrating which country each athlete represents.

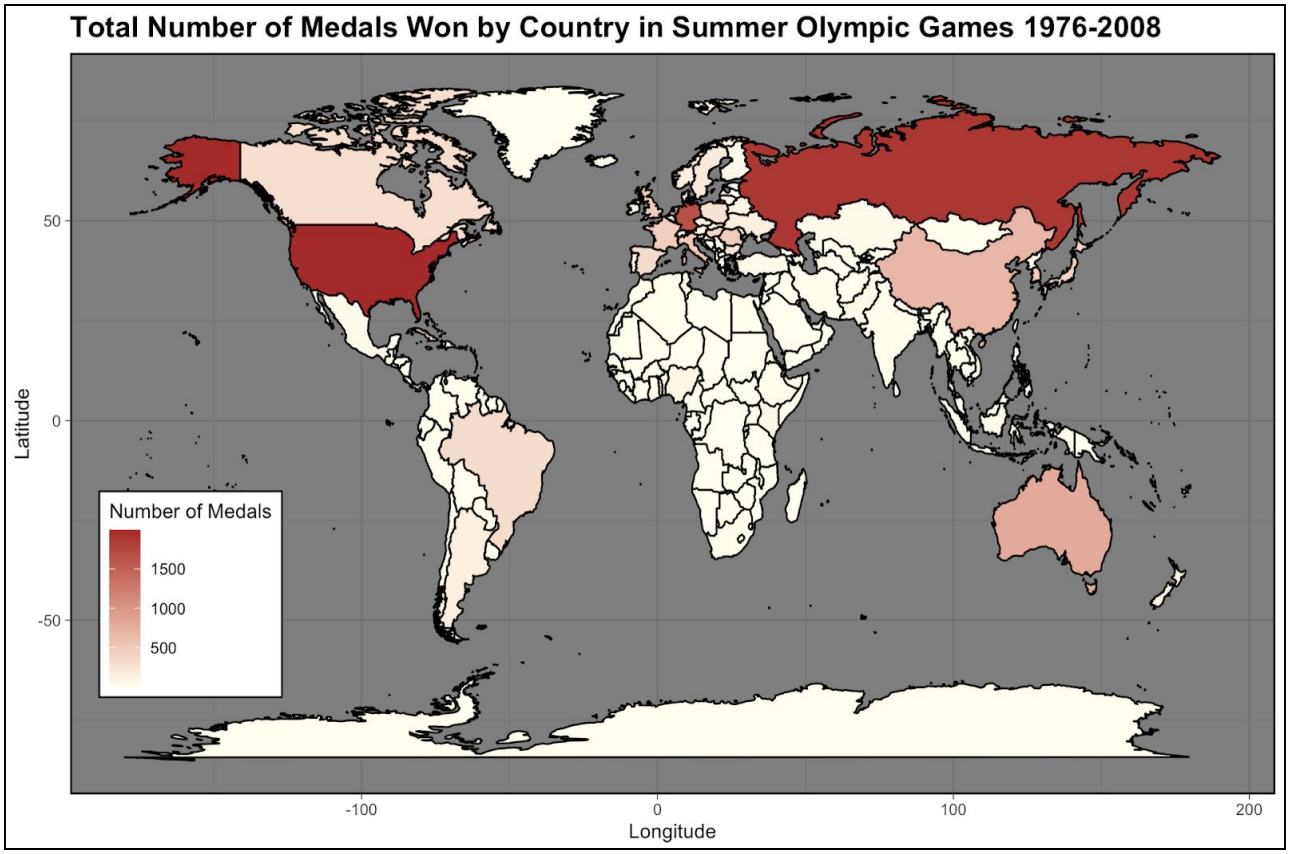

Figure 3: Choropleth of Total Number of Medals Won per Country

The choropleth in Figure 3 illustrates geographically the total number of medals won for each country in the Summer Olympics from 1976 to 2008. From this visualization, it is overwhelmingly apparent that two of the most successful countries in the Summer Olympics over the course of this 32-year timespan were the United States and Russia (winning nearly 2,000 medals apiece). Germany too, appears to have won a substantial number of medals. Australia, China as well as many of the European nations also look to have done considerably well in terms of total medal counts. Nonetheless, the emergent theme of this particular graph seems to be the dominant success of the United States and Russia (as well as Germany).

This visualization was created by joining the world map data (with latitude and longitude coordinates) in R to the Summer Olympic Medals dataset by country. A continuous color scale from ivory to brown was selected. The color was mapped to the total number of medals won per country. Initially, there were a few countries with missing data. However, through the drafting process these were set to a default color of ivory, in order to preserve the overall effect of the visualization. Finally, the legend was repositioned to the lower left corner of the graph.

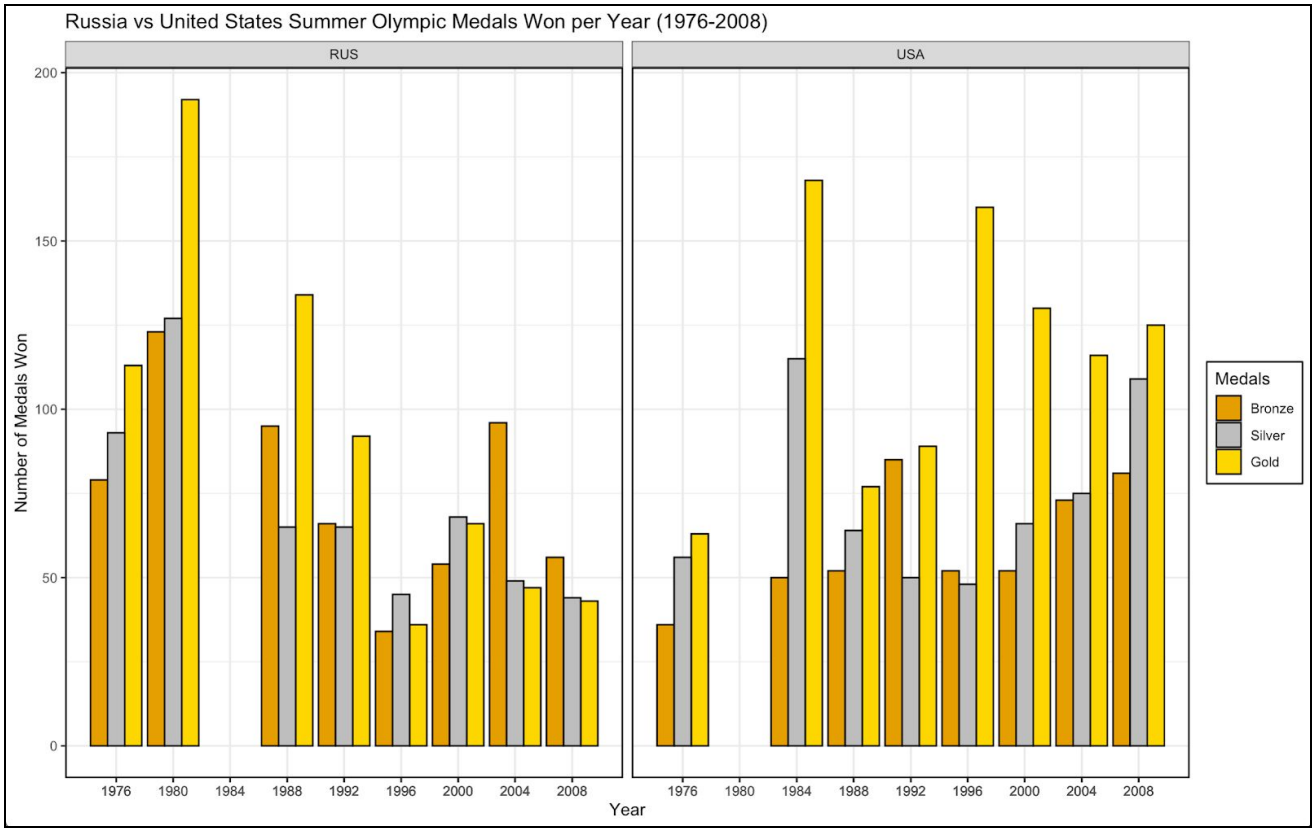

Figure 4: Small Multiples Dodged Bar Graph of Gold, Silver & Bronze Medals Won per Year for US and Russia

The small multiples dodged bar chart shown in Figure 4 compares the individual number of bronze, silver, and gold medals won by Russia and the United States in each Summer Olympic Games from 1976 to 2008. Since the United States and Russia were the two most successful nations in terms of overall medals won at the Summer Olympics during the timeframe under analysis, comparing the individual medal counts of these two countries over time was warranted. From this visualization, it is apparent that the United States won more gold medals than any other medal type in every single Summer Olympic year, while Russia's top winning medal types are a bit more mixed. Additionally, it is evident that Russia's best Summer Olympic year in terms of medals won was 1980. Coincidentally, this also happens to be the year that the United States boycotted the Summer Olympic Games. Likewise, Russia did not compete in the following Summer Olympics in 1984.

This visualization was constructed in R by using a facet wrap to display the two countries side-by-side for comparison purposes. The individual medals were dodged to display the three different medal types (bronze, silver, and gold) alongside one another instead of being stacked. Finally, color was again appropriately mapped to the specific medal type. Through the drafting process, it was discovered that using a facet wrap was more effective than a facet grid for comparison purposes of the two countries in this particular small multiples graph.

Animated Visualization:

Animation

The animated visualization at the above link shows the total medals won over time for the top six countries in the Summer Olympic Games from 1976 to 2008. On the left is a bar chart which aggregates the number of medals won for each nation over time, and on the right is a line plot that shows the cumulative medals won over time for these same six countries. From this visualization, it is evident that the United States, Russia, and Germany were very competitive with one another, outperforming all other nations by a considerable margin. The three next best countries in terms of total medals won over time were Australia, China, and Italy. This visualization also illustrates that China did not participate in either the 1976 or 1980 Summer Olympics. Consequently, their medal tally begins later than the five other countries shown.

This visualization was constructed in R markdown by first creating a rolling total of all medals won for each of the top six countries over the course of nine Summer Olympic Games. Both the animated bar chart and animated line plot used a discrete pastel color palette from the color brewer library that was mapped to the individual countries, and a dark theme/background to tie the two graphs together visually. The axes of the bar chart were flipped in order to allow for the country names to be more easily read along the y-axis, and the country names were added at the end of each line in the line graph via the direct labels library. Finally, both animations were saved as separate gifs, and then iteratively stitched together, frame by frame, in order to create a new, combined gif displaying both animated plots side-by-side as a single visualization.

There were a number of interesting discoveries made from this data. First, it was discovered that Athletics, Aquatics, and Rowing were the three most popular sports in terms of the total number of medalists at the Summer Olympics from 1976 to 2008. This was illustrated through the use of a treemap in Figure 1. I also learned that Michael Phelps was the most successful Summer Olympic athlete during the timeframe under analysis, winning 14 gold, and 2 bronze medals. This was shown through a stacked bar graph in Figure 2. Next, it was determined that the United States, Russia, and Germany were the most dominant nations in terms of total medals won at the Summer Olympics between 1976 and 2008. This was displayed via a choropleth in Figure 3. Finally, I explored the comparison of the individual number of gold, silver, and bronze medals won for the top two countries, the United States and Russia, through the use of a small multiples dodged bar graph in Figure 4. This visualization illustrated that the United States won more gold medals than any other medal type in every single Summer

Olympic year, while Russia's results were a bit more mixed. It was also discovered that Russia's most successful Summer Olympic year happened to coincide with the same year that the United States boycotted the Summer Olympic Games. Additionally, an animated plot of the top six countries at the Summer Olympics from 1976 to 2008 illustrated that the United States, Russia, and Germany outperformed all other nations in terms of total medals won by a considerable margin.

Given more time, it would have been very insightful to evaluate the population of each country over time, and how this relates to the total number of medals won for each respective nation. The dataset did not contain this information. However, Germany for example has a current population of just about 84 million people, while the United States' population is approximately 331 million. Consequently, it is very impressive that Germany remained so competitive with the United States in terms of total medal counts in the Summer Olympics between 1976 and 2008 given that their population is only a small fraction of ours.

Appendix

R Code Used

Figure 1:

```
library(ggplot2)
library(treemapify)
library(treemap)
library(dplyr)

data <- group_by(olympics_II, Sport, Discipline) %>% summarize(Medalists =
sum(Bronze)+sum(Silver)+sum(Gold))

ggplot(data, aes(area=Medalists, fill= Medalists, label=Sport, subgroup=Discipline))+
  geom_treemap()+
  geom_treemap_subgroup_text(place="centre", grow=T, alpha=.9, colour="White",
fontface="italic", min.size=0)+
  geom_treemap_text(colour="Red", place="topleft", reflow=T)+
  labs(fill="Number of Medalists")+
  ggtitle("Largest Sports by Discipline in Summer Olympic Games 1976-2008")+
  theme(plot.title = element_text(size = 16, face="bold"))+
  theme(panel.border=element_rect(colour="black", fill=NA, size=1),
        legend.background = element_rect(colour="black", size=0.5))+
  scale_fill_continuous(trans = 'reverse', guide = guide_colourbar(reverse=T))
```

Figure 2:

```
library(ggplot2)
library(dplyr)
library(ggflags)
olympic_new <- olympics_II

phelps <- filter(olympic_new, Athlete == "PHELPS, Michael")
andrianov <- filter(olympic_new, Athlete == "ANDRIANOV, Nikolay")
fischer <- filter(olympic_new, Athlete == "FISCHER, Birgit")
nemov <- filter(olympic_new, Athlete == "NEMOV, Alexei")
```

```
thompson <- filter(olympic_new, Athlete == "THOMPSON, Jenny")
torres <- filter(olympic_new, Athlete == "TORRES, Dara")
biondi <- filter(olympic_new, Athlete == "BIONDI, Matthew")
coughlin <- filter(olympic_new, Athlete == "COUGHLIN, Natalie")
dittyatin <- filter(olympic_new, Athlete == "DITYATIN, Aleksandr")
hall <- filter(olympic_new, Athlete == "HALL, Gary Jr.")

ten <- rbind(phelps, andrianov, fischer, nemov, thompson, torres, biondi, coughlin, dittyatin, hall)

athlete <- group_by(ten, Athlete, Country_Code, Medal) %>% summarize(All_Medals =
sum(Bronze)+sum(Silver) + sum(Gold))

athlete$Country_Code[which(athlete$Country_Code == "USA")] = "us"
athlete$Country_Code[which(athlete$Country_Code == "RUS")] = "ru"
athlete$Country_Code[which(athlete$Country_Code == "GER")] = "de"

athlete$Medal <- factor(athlete$Medal, levels = c("Bronze", "Silver", "Gold"))
athlete$test <- paste(athlete$Athlete, '     ')

ggplot(athlete, aes(x = reorder(test, All_Medals), y = All_Medals, fill=Medal)) +
  geom_bar(stat="identity", colour= "black", size= .3, width= .4)+
  geom_flag(y = -1.1, aes(country = Country_Code), size = 5) +
  coord_flip() +
  scale_fill_manual(values = c("#E69F00", "grey", "gold"), guide = guide_legend(reverse = T)) +
  scale_y_continuous(breaks=c(0,2,4,6,8,10,12,14,16))+
  labs(y="Number of Medals", x="Athlete")+
  ggtitle("Top Ten Athletes with Most Summer Olympic Medals, 1976-2008")+
  theme_bw()+
  theme(panel.border=element_rect(colour="black", fill=NA, size=1),
        legend.background = element_rect(colour="black", size=0.5),
        axis.ticks.y=element_blank(),
        panel.grid.minor = element_line(size = 0.25, linetype = 'dashed', colour = "grey"),
        panel.grid.major = element_line(size = 0.35, linetype = 'dashed', colour = "grey"))
```

Figure 3:

```
library(ggplot2)
library(mapproj)
library(dplyr)

world = map_data('world')
ol <- olympics_II

ol <- group_by(ol, Country) %>% summarize(All_Medals = sum(Bronze)+sum(Silver) +
sum(Gold))

ol$Country[which(ol$Country == "United States")] = "USA"
ol$Country[which(ol$Country == "Czechoslovakia")] = "Czech Republic"
ol$Country[which(ol$Country == "United Kingdom")] = "UK"
ol$Country[which(ol$Country == "Korea, North")] = "North Korea"
ol$Country[which(ol$Country == "Korea, South")] = "South Korea"

olympic_map <- left_join(world, ol, by = c("region" = "Country"))

ggplot(olympic_map,
      aes(x=long, y=lat, group=group, fill= All_Medals))+
  geom_polygon(colour = "black")+
  scale_fill_continuous(na.value = "ivory", low = "ivory", high = "brown")+
  labs(x="Longitude",
      y="Latitude",
      fill="Number of Medals") +
  ggtitle("Total Number of Medals Won by Country in Summer Olympic Games 1976-2008")+
  theme_dark()+
  theme(plot.title = element_text(size = 16, face="bold"),
        panel.border=element_rect(colour="black", fill=NA, size=1),
        legend.background = element_rect(colour="black", size=0.45),
        legend.position= c(0.1, 0.27))
```

Figure 4:

```
library(dplyr)
library(ggplot2)
USA <- filter(olympics_II, Country_Code == "USA")
RUS <- filter(olympics_II, Country_Code == "RUS")

USA <- group_by(USA, Year, Country_Code, Medal, All) %>% summarize(All_Medals =
sum(Bronze)+sum(Silver) + sum(Gold))

RUS <- group_by(RUS, Year, Country_Code, Medal, All) %>% summarize(All_Medals =
sum(Bronze)+sum(Silver) + sum(Gold))

USA_RUS <- rbind(USA, RUS)

USA_RUS$Medal <- factor(USA_RUS$Medal, levels = c("Bronze", "Silver", "Gold"))

USA_RUSbar <- ggplot(USA_RUS, aes(x=Year, y=All_Medals, fill=Medal))+
  geom_col(position = "dodge", colour = "black")

USA_RUSbar + facet_wrap(Country_Code ~ .) +
  scale_x_continuous(breaks=USA_RUS$Year) +
  labs(x="Year",
       y="Number of Medals Won",
       fill="Medals") +
  scale_fill_manual(values = c("#E69F00", "grey", "gold")) +
  ggtitle("Russia vs United States Summer Olympic Medals Won per Year (1976-2008)") +
  theme_bw()+
  theme(panel.border=element_rect(colour="black", fill=NA, size=1),
        legend.background = element_rect(colour="black", size=0.5),
        panel.grid.minor.x = element_blank())
```

Animation (R Markdown):

```
---
title: "<span style='font-size: 15px'>Total Medals Won Over Time For Top Six Countries in Summer Olympic Games (1976-2008)</style>"
output:
  html_document: default
---
```

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = FALSE, warning=FALSE)
knitr::opts_chunk$set(fig.width=8, fig.height=6)
```

```{r, echo=FALSE, warning=FALSE, message=FALSE}
setwd("/Users/jonathanlynch/Desktop/DSC 465")
olympics_V <- read.csv("olympics_V.csv", header = TRUE, sep = ",")
library(ggplot2)
library(gganimate)
library(gifski)
library(knitr)
library(dplyr)
library(magick)
library(directlabels)

u <- filter(olympics_V, Country_Code == "USA")
u <- group_by(u, Year, Country_Code) %>% summarize(All_Medals =
sum(Bronze)+sum(Silver) + sum(Gold))
u[, 3] <- cumsum(u[, 3])

r <- filter(olympics_V, Country_Code == "RUS")
r <- group_by(r, Year, Country_Code) %>% summarize(All_Medals = sum(Bronze)+sum(Silver)
+ sum(Gold))
r[, 3] <- cumsum(r[, 3])

a <- filter(olympics_V, Country_Code == "AUS")
a <- group_by(a, Year, Country_Code) %>% summarize(All_Medals =
sum(Bronze)+sum(Silver) + sum(Gold))
a[, 3] <- cumsum(a[, 3])
```

```
g <- filter(olympics_V, Country_Code == "GER")
g <- group_by(g, Year, Country_Code) %>% summarize(All_Medals =
sum(Bronze)+sum(Silver) + sum(Gold))
g[, 3] <- cumsum(g[, 3])

c <- filter(olympics_V, Country_Code == "CHN")
c <- group_by(c, Year, Country_Code) %>% summarize(All_Medals = sum(Bronze)+sum(Silver)
+ sum(Gold))
c[, 3] <- cumsum(c[, 3])

i <- filter(olympics_V, Country_Code == "ITA")
i <- group_by(i, Year, Country_Code) %>% summarize(All_Medals = sum(Bronze)+sum(Silver)
+ sum(Gold))
i[, 3] <- cumsum(i[, 3])

top_six <- rbind(u, r, a, g, c, i)

ggplot(top_six, aes(x = reorder(Country_Code, -All_Medals), y = All_Medals,
fill=Country_Code))+
  geom_bar(stat="identity", colour="black", size= .3, show.legend= FALSE) +
scale_fill_brewer(palette="Pastel1") +
  coord_flip() + transition_time(Year) +
  theme_dark()+
  labs(title = "Year: {frame_time}", x="Country", y="Medals Won")

anim_save("bar.gif")

ggplot(top_six, aes(x=Year, y=All_Medals, colour=Country_Code))+
  geom_point(size=.9)+
  geom_line(size=.9)+
  scale_color_brewer(palette = "Pastel1", guide = 'none') +
  scale_x_discrete(limits=c(1976,1980,1984,1988,1992,1996,2000,2004,2008)) +
  geom_dl(aes(label = Country_Code), method = list(dl.trans(x = x + 0.2), "last.points", cex = .8,
fontface="bold")) +
  theme_dark()+
```

```
  transition_reveal(Year)+
  labs(y="Medals Won", title="")

anim_save("line.gif")

bar_gif <- image_read("bar.gif", strip = TRUE)
line_gif <- image_read("line.gif", strip = TRUE)

new_gif <- image_append(c(bar_gif[1], line_gif[1]))
for(i in 2:100){
  combined <- image_append(c(bar_gif[i], line_gif[i]))
  new_gif <- c(new_gif, combined)
}

new_gif
```