FinalProjEric

Eric Farrall

2025-05-07

1 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 1 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 occuring 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 1. However, there is a general trend of an increase of appearances over time for all countries, which can be seen more clearly in Figure 2, which has gaps removed for clarity.
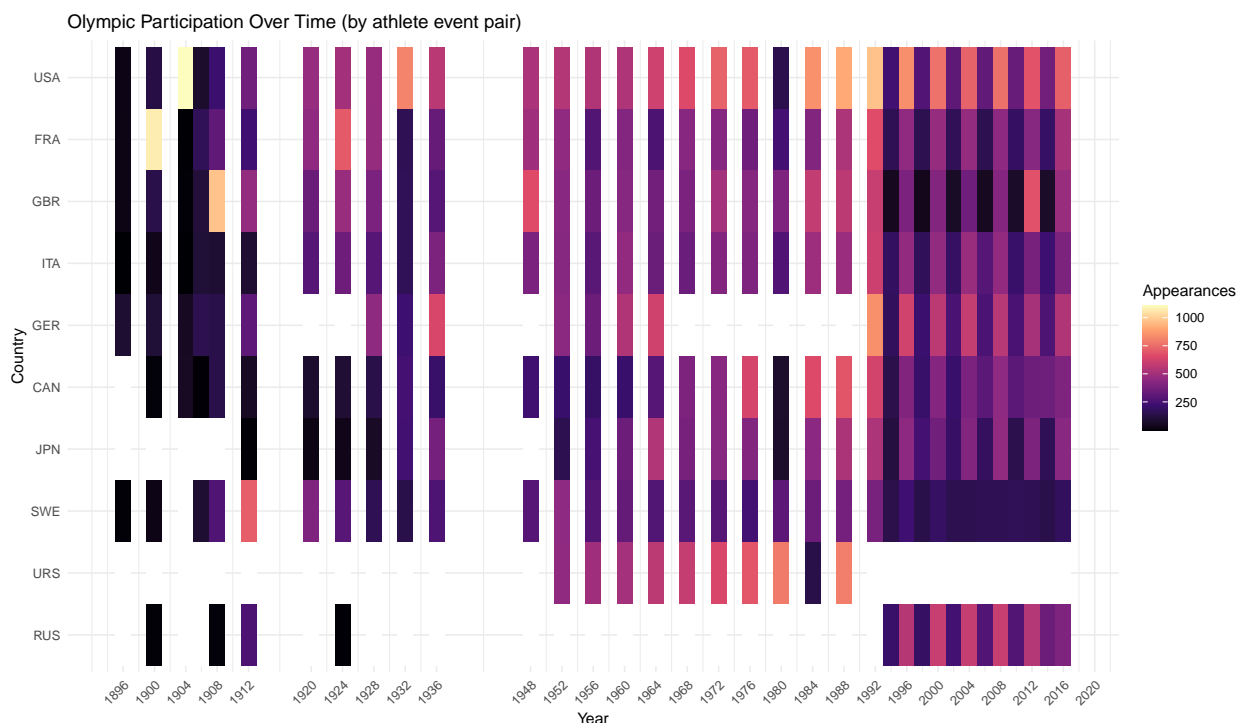
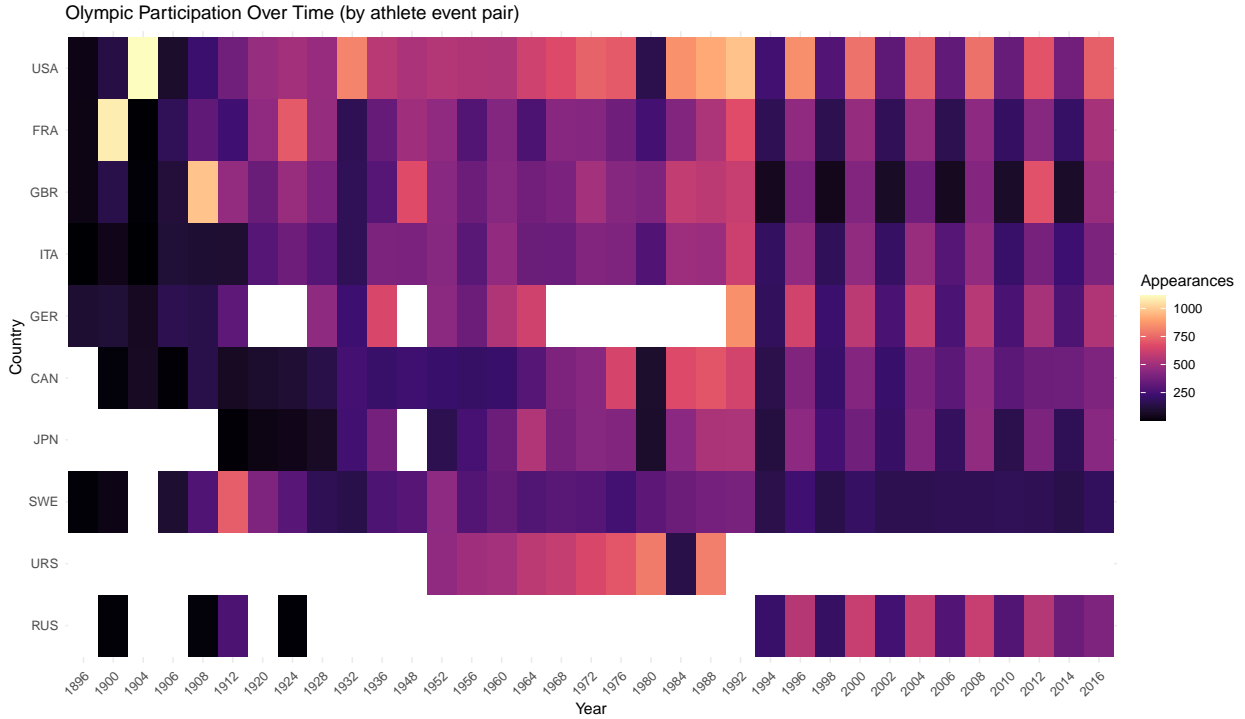Figure 1: Olympic Participation - Gaps

Figure 2: Olympic Participation - No Gaps

Olympic Participation Over Time (by athlete event pair)

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 2, 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 4 and ranked 7th in Figure 3. 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 definately do specialize in a particular games. This is apparant in the Canadian example, whose cold climite encourages them to pursue winter games.

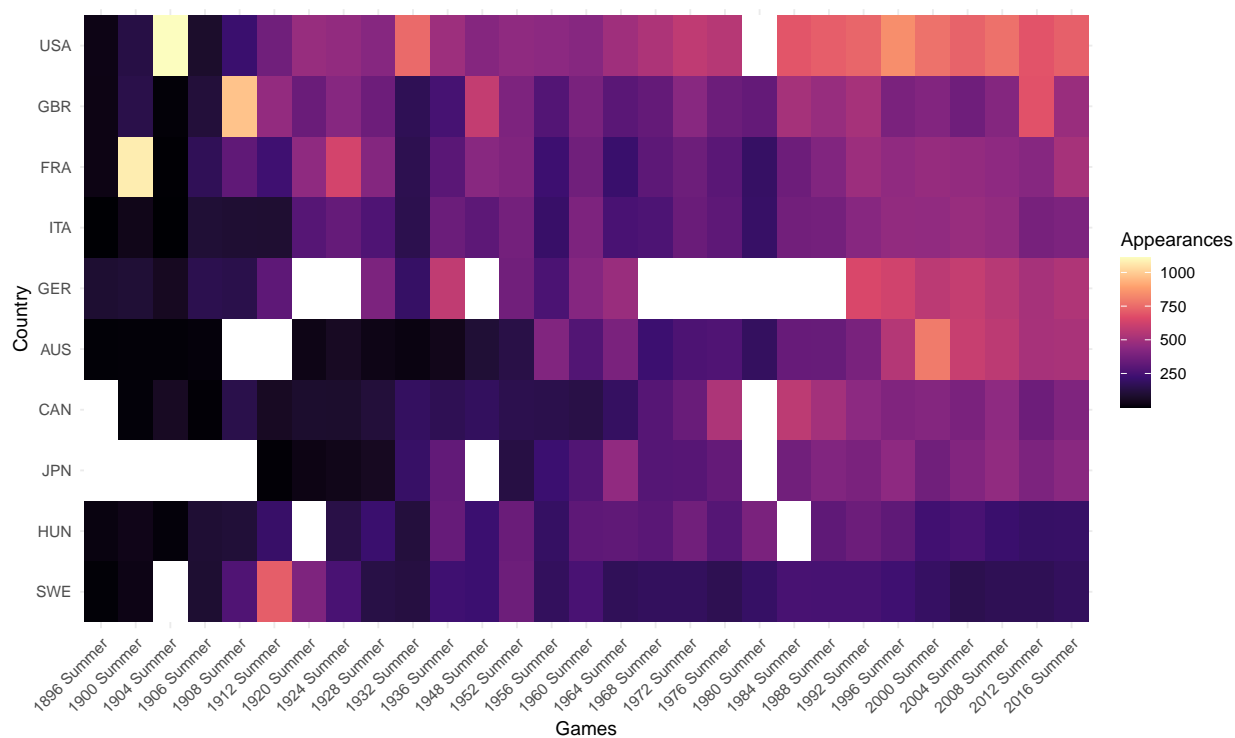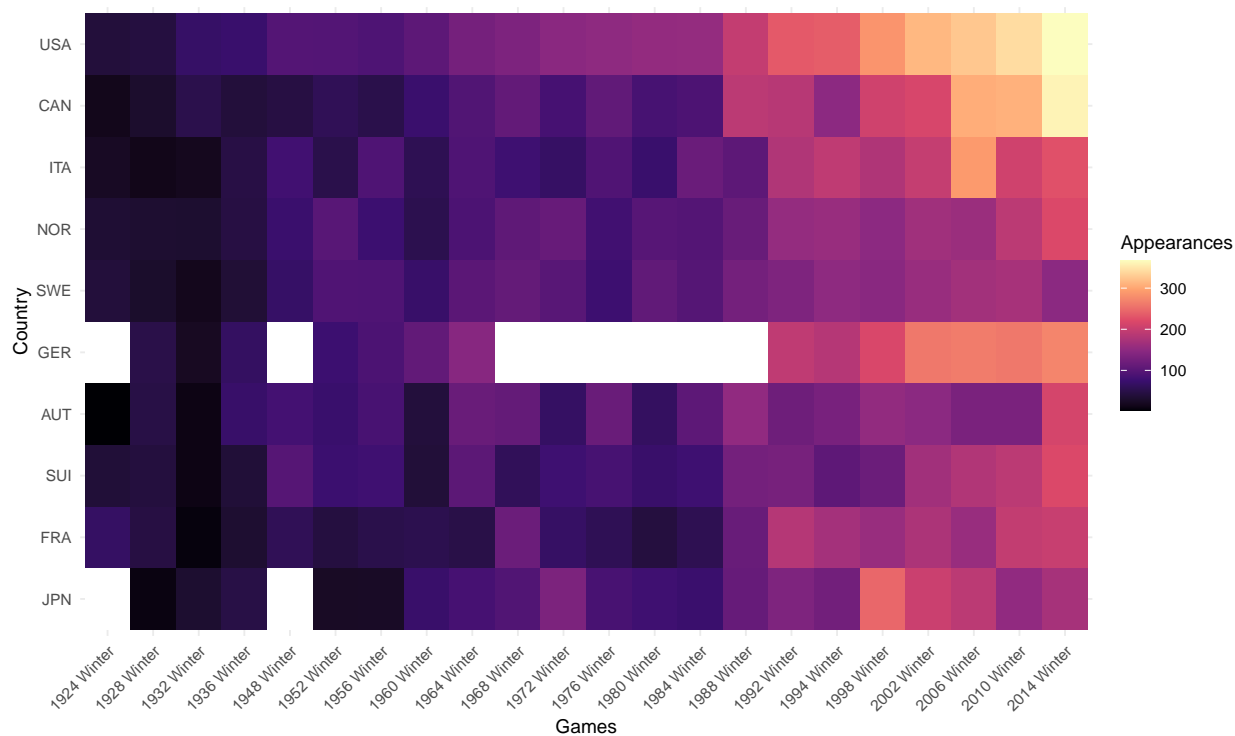Figure 3: Summer Olympic Participation Over Time (by athlete event pair)

Figure 4: Winter Olympic Participation Over Time (by athlete event pair)

```r
library(tidyverse)
library(ggplot2)
library(scales)

url <- "https://huggingface.co/datasets/EFarrallpsu/STAT184_Eric_Jackson_Nina/resolve/main/athl
athletes <- read.csv(url)

#transform into dataframe with case as year, with columns for each country signifying appearan
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 differentation later on

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



# 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 contries 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)

#plot showing total appearances over time in top 10 countries
```

```r
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
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))

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

View(season_case)
# converts vector of strings signifying number of appearances to integer value
for (col in names(season_case)[-1]) {
```

```r
    season_case[[col]] <- sapply(season_case[[col]], function(x) if (is.null(x)) NA else length(
}


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


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))
#plot for summer games
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
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))
```