Lab 1

Student names here

Lab Overview

This lab will explore basic `ggplot2` functionality. Please turn in one compiled document (PDF) per group. We will use two datasets related to the recent XXXIII Olympiad (Olympics) held in Paris, France.

Medal Count Dataset

The first dataset contains medal counts for all countries earning a medal.

```r
library(tidyverse)
medals <- read_csv('https://raw.githubusercontent.com/stat408/Data/main/MedalCount.csv')
```

1. (4 points)

Create a figure that tells the story of the medal count at the Paris Olympics.

```
medals |>
  filter(Gold > 0) |>
  ggplot(aes(y = Country, x = `Total Medals`)) +
  geom_point() +
  labs(title = 'Total medal count from 2024 Olympics',
       subtitle = 'Countries with at least one gold medal',
       caption = 'source = https://www.kaggle.com/datasets/berkayalan/paris-2024-olympics-med
  theme_minimal()
```

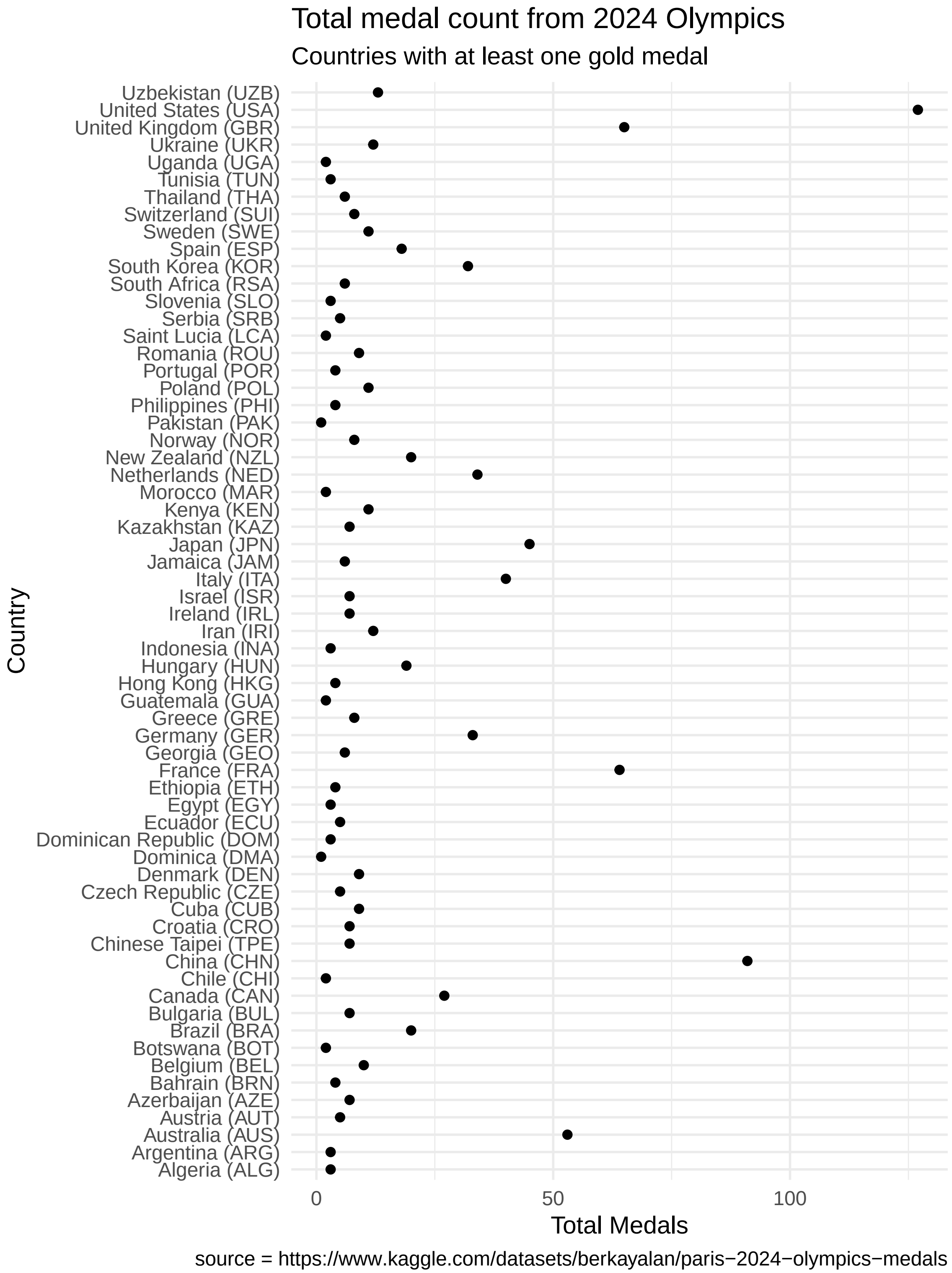


Figure 1: A figure summarizing the medal count for countries from the 2024 Paris Olympics. Only countries with at least one gold medal are displayed.

Olympic Athlete Dataset

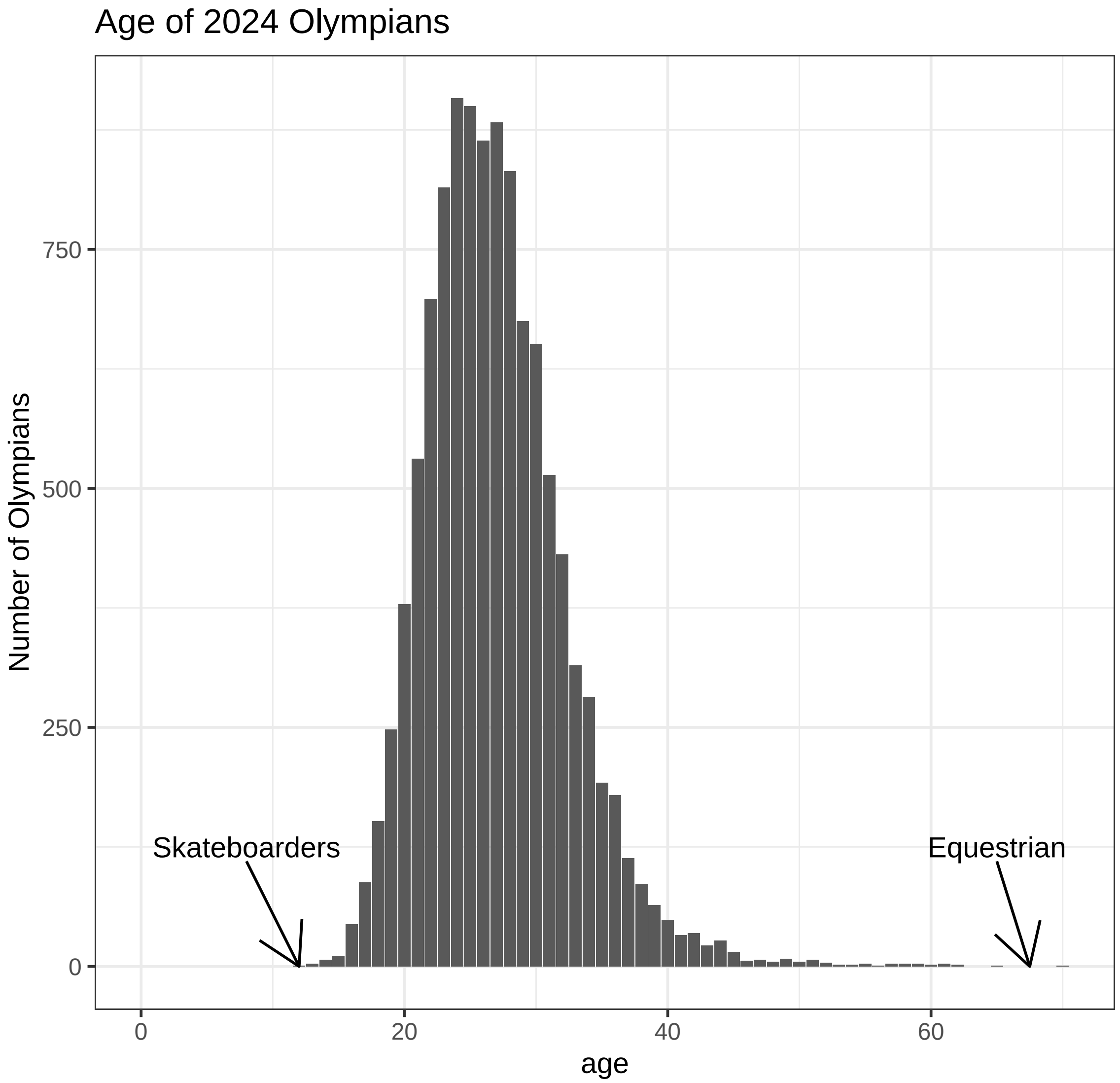
This set of figures will use an Olympic dataset from Kaggle. Additional information is available at https://www.kaggle.com/datasets/willianoliveiragibin/olympics-2024?resource=download&select=athletes+new.csv

```
athletes <- read_csv('https://raw.githubusercontent.com/stat408/Data/main/athletes%20new.csv
  mutate(birth_year = year(birth_date)) # extracts birth year
```

2. (4 points)

Using the `birth_date` variable, create a figure that visualizes the ages of the Olympians. Which sports tend to have the youngest and oldest athletes?

```
athletes |>
  mutate(age = 2024 - birth_year) |>
  ggplot(aes(x = age)) +
  geom_bar() +
  xlim(0,NA) +
  theme_bw() +
  annotate('text', x = 8, y = 125, label = 'Skateboarders') +
  annotate("segment", x=8, y=110, xend=12, yend=0, arrow = arrow()) +
  annotate('text', x = 65, y = 125, label = 'Equestrian') +
  annotate("segment", x=65, y=110, xend=67.5, yend=0, arrow = arrow()) +
  ylab("Number of Olympians") +
  ggtitle('Age of 2024 Olympians') +
   labs(caption = 'source = https://www.kaggle.com/datasets/willianoliveiragibin/olympics-20
```



source = https://www.kaggle.com/datasets/willianoliveiragibin/olympics−2024?resource

Figure 2: This figure shows the age of the 2024 Olympians. Note the youngest athletes are skateboarders and the oldest athletes are equestrians - that riders not the horses.

3. (4 points)

Create a figure that displays the number of competing athletes from the 12 countries with the most medals.

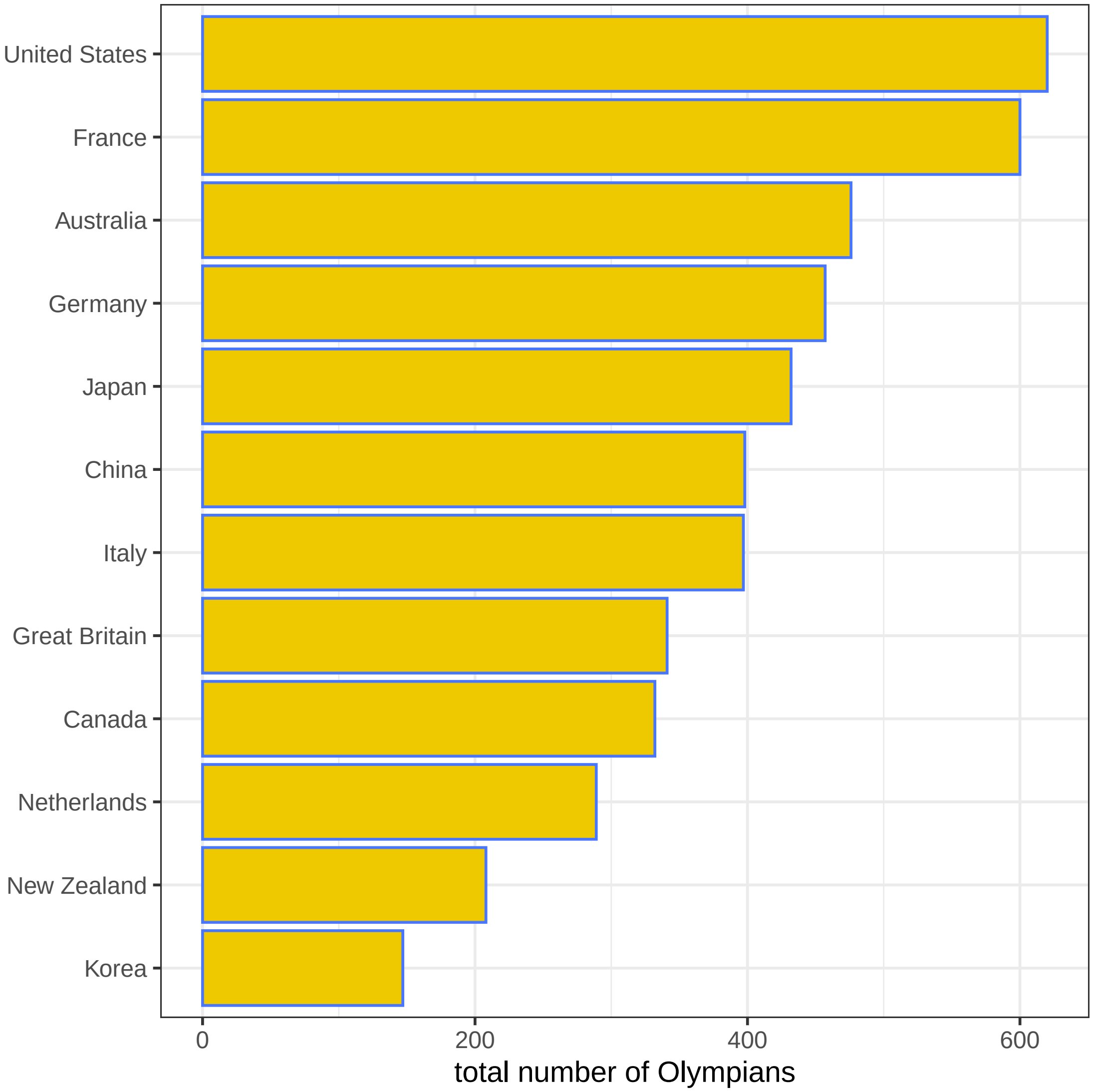
```r
# function to extract country abbreviation
extract_country_abbr <- function(string_in){
  str_split(string_in, '\\(' )[[1]][2] |>
  str_sub(end = -2)
}

top12 <- sapply(medals$Country[1:12], extract_country_abbr)

top12_athletes <- athletes |>
  filter(country_code %in% top12)
```

Note: I've made this process easier for you by only including athletes from these 12 countries (USA, CHN, JPN, AUS, FRA, NED, GBR, KOR, ITA, GER, NZL, CAN).

```r
top12_athletes |>
  group_by(country) |>
  tally() |>
  ggplot(aes(y=reorder(country, +n), x = n)) +
  geom_bar(stat="identity", color = 'royalblue1', fill = 'gold2') +
  ylab('') +
  xlab('total number of Olympians') +
  theme_bw() +
   labs(caption = 'source = https://www.kaggle.com/datasets/willianoliveiragibin/olympics-20
```



source = https://www.kaggle.com/datasets/willianoliveiragibin/olympics−2024?resource

Figure 3: Total Olympians from the 12 countries earning the most medals at the 2024 Olympics.

4. (4 points)

Use the `Q4_data` to visualize the relationship between the number of medals earned by a country against the number of athletes participating in the Olympics.

```r
medals$country_code <- sapply(medals$Country, extract_country_abbr)

Q4_data <- athletes |>
  group_by(country_code, country_full) |>
  tally() |>
  rename(num_athletes = n) |>
  left_join(medals, by = 'country_code') |>
  mutate(`Total Medals` = case_when(
    !is.na(`Total Medals`) ~ `Total Medals`,
    is.na(`Total Medals`) ~ 0
  )) |>
  select(country_code, num_athletes, country_full, `Total Medals`) |>
  rename(total_medals = `Total Medals`)
```

```r
Q4_data |>
  ggplot(aes(y = total_medals, x = num_athletes)) +
  geom_point(data = Q4_data |> filter(total_medals < 25)) +
  geom_smooth(method = 'loess', formula = 'y ~ x', se = F) +
  ylab('Total Medals') +
  xlab('Number of Olympians') +
  geom_text(data = Q4_data |> filter(total_medals > 25), aes(label = country_code)) +
  theme_bw()
```

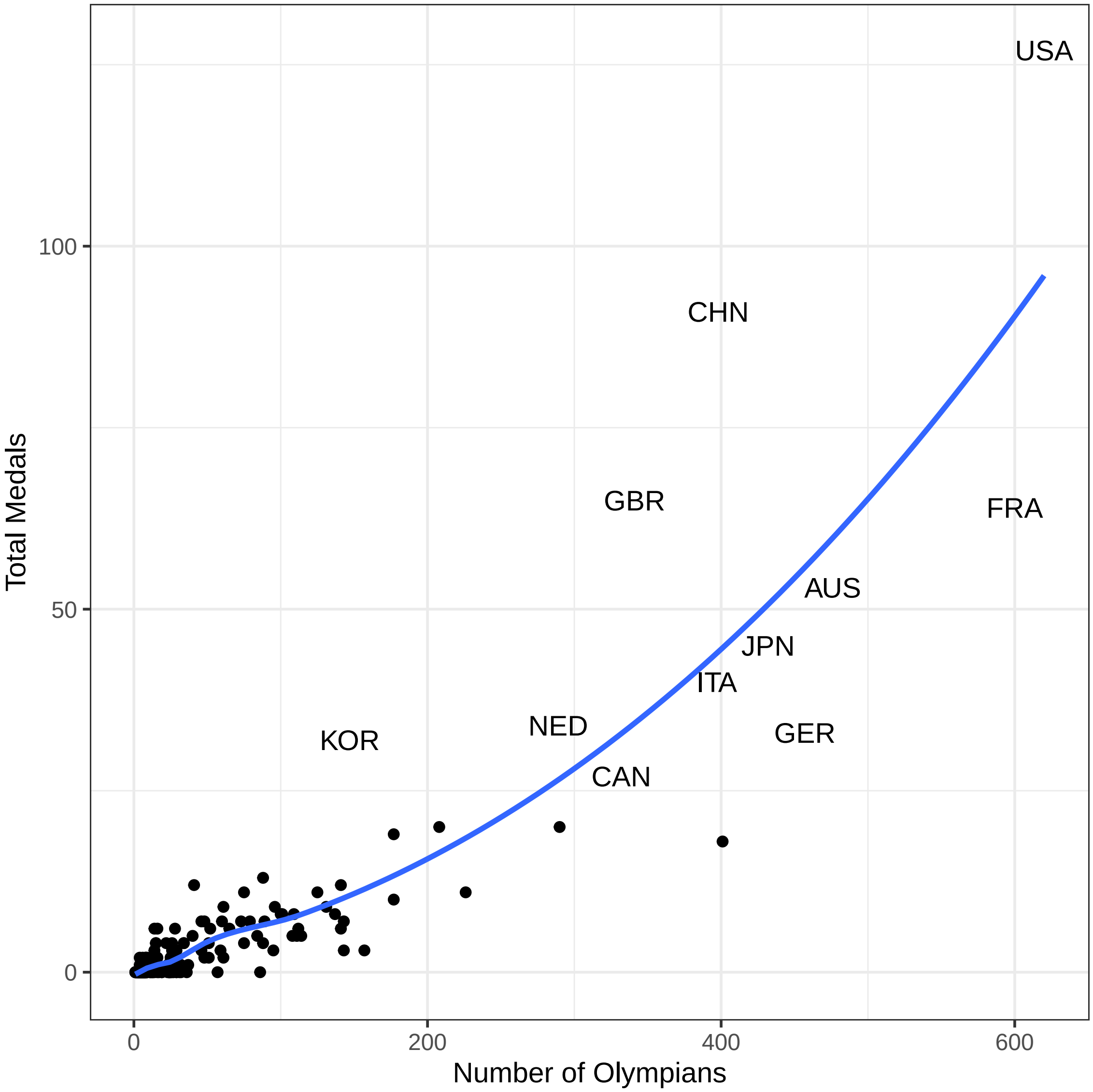


Figure 4: This figure displays the relationship between the total number of Olympians against medals earned by country. Countries earning more than 25 medals have the are indicated by the country code. Countries above the line tend to overperform and countries under the line tended to underperform.

5. (4 points)

The `athletes` dataset also contains the events the athletes are competing in. See the value for Montana's Katherine Berkoff

[“Women’s 100m Backstroke”, “Women’s 4 x 100m Medley Relay”]

or gymnast Simone Biles

[‘Women’, “Women’s All-Around”, “Women’s Balance Beam”, “Women’s Floor Exercise”, “Women’s Uneven Bars”, “Women’s Vault”, “Women’s Team”]

Describe (in words or pseudocode) what you’d need to do and/or what additional information you’d need in order to create a figure that displayed the number of events competed in by athletes that won medals.