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.

```
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.

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

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

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?

3. (4 points)

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

```
string_in <- medals$Country[1]

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).

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.

```
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`)
```

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.