Project 1

This is the dataset you will be working with:

```
olympics <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience
/tidytuesday/master/data/2021/2021-07-27/olympics.csv')

triathlon <- olympics |>
  filter(!is.na(height)) |>             # only keep athletes with known height
  filter(sport == "Triathlon") |>       # keep only triathletes
  mutate(
    medalist = case_when(                # add column to track medalist vs not
      is.na(medal) ~ "non-medalist",
       !is.na(medal) ~ "medalist"          # any medals (Gold, Silver, Bronze)
count
    )
  )
```

`triathlon` is a subset of `olympics` and contains only the data for triathletes. More information about the original `olympics` dataset can be found on the tidytuesday project and on Olympedia.

For this project, use `triathlon` to answer the following questions about athletes competing in this sport:

1. In how many events total did male and female triathletes compete for each country?
2. Are there height differences among triathletes between sexes or over time?
3. Are there height differences among triathletes that have medaled or not, again also considering athlete sex?

You should make one plot per question.

Hints:

- We recommend you use a bar plot for question 1, a boxplot for question 2, and a sina plot overlaid on top of violins for question 3. However, you are free to use any of the plots we have discussed in class so far.
- For question 2, you will have to convert `year` into a factor.
- For question 3, consider why a boxplot or simple violin plot is not a good idea and mention this in the approach section.
- For all questions, you can use either faceting or color coding or both. Pick whichever you prefer.
- Adjust `fig-width`, `fig-height`, and `out-width` in the chunk options to customize figure sizing and figure aspect ratios. `fig-width` and `fig-height` are given in inches and will usually be between 3 and 10. `out-width` is given in percent and will usually be between 50% and 100%.

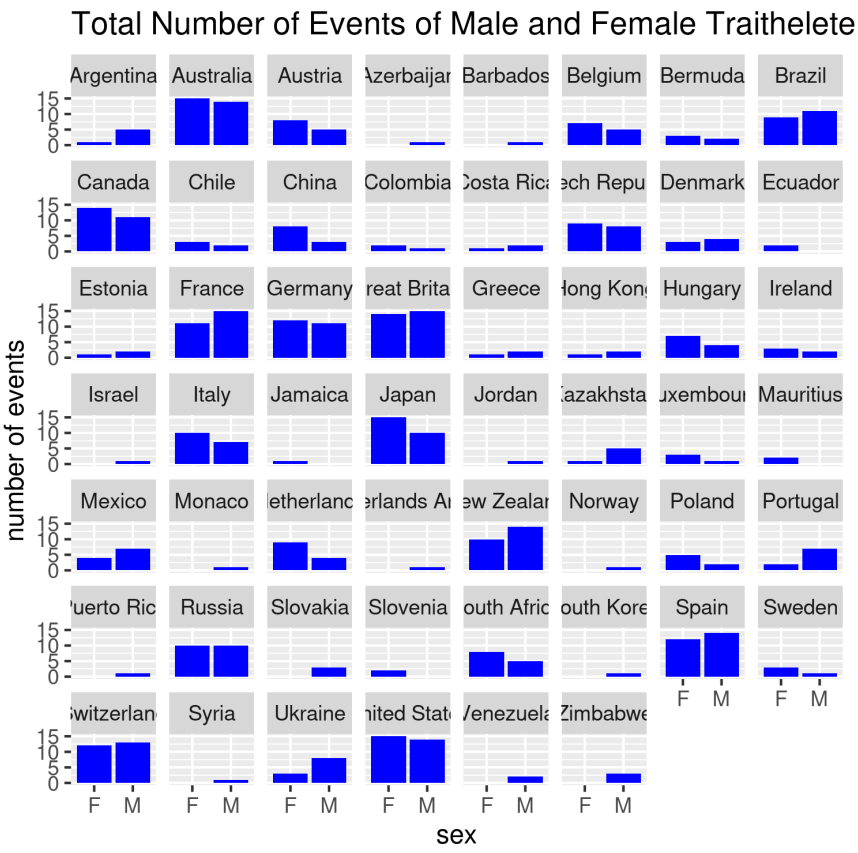
You can delete these instructions from your project. Please also delete text such as *Your approach here* or `# Q1: Your R code here`.

Introduction: *We are working with the `triathlon` dataset, which contains 528 athlete records from the Modern Olympic games from Athens in 1896 to Rio 2016 that participated in Triathlon events. In this dataset, each row corresponds to one athlete and there are 15 columns providing information about the athlete with information about the Olympic event they competed in. Information about the athlete is shown in the id, name, sex, age, height, weight, team, noc. Information about the associated Olympic game they competed in is shown in games, year, season, city, sport, event, medal, and medalist column. To answer the three questions about, we will work with five variables, the athletes sex(column sex), the athletes country(column team), the athletes height (column height), and the athletes medal status (column medalist), and the year of the event the athlete competed in (year). The sex is provided as a categorical value as either a F or M. The athletes country is provided as a categorical value as the name of their country. The athletes height is provided as a numeric value in centimeters The year of the event the atheltes competed in provided as a numerical value by year number.*

Approach: *To show the total number of events that male and female triathletes competed for each country we will use a simple bar plot geom_bar(). To show the differences of height between the sexes of triatheletes over time we will be using a boxplot geom_boxplot(). We will also be converting year into a factor to do our boxplot. Finally to show if there are height differences among triathletes that have medaled or not keeping in mind their sex we will use a sina plot geom_sina() overlaid on violin plots geom_violin(). This a good option to answer this question as violins make it easier to compare multiple distributions side-by-side.*

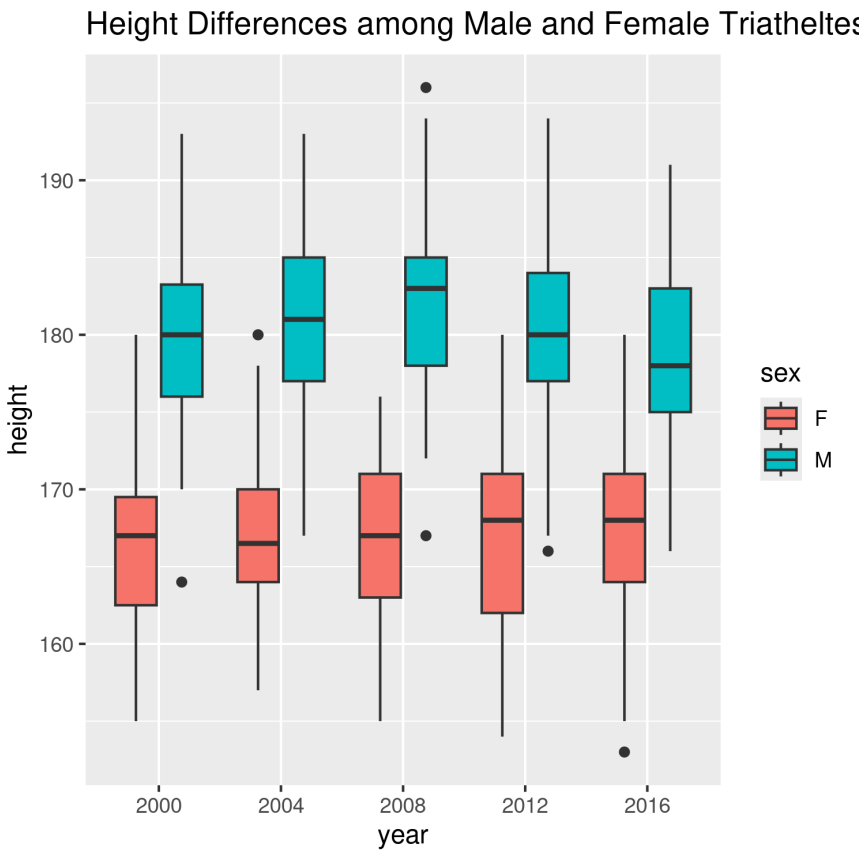
Analysis: Question 1: In how many events total did male and female triathletes compete for each country? To answer this question, we make a simple bar plot of the number of events by sex per country

```
# Use facetwrap to display each female and male events per each country and labs
to label the plot


ggplot(triathlon, aes(sex))+
  geom_bar(position = "dodge", fill = "blue")+
  facet_wrap(~team, dir = "h") +
  labs(
   title = "Total Number of Events of Male and Female Traitheletes per Country",
    y = "number of events"
  )
```



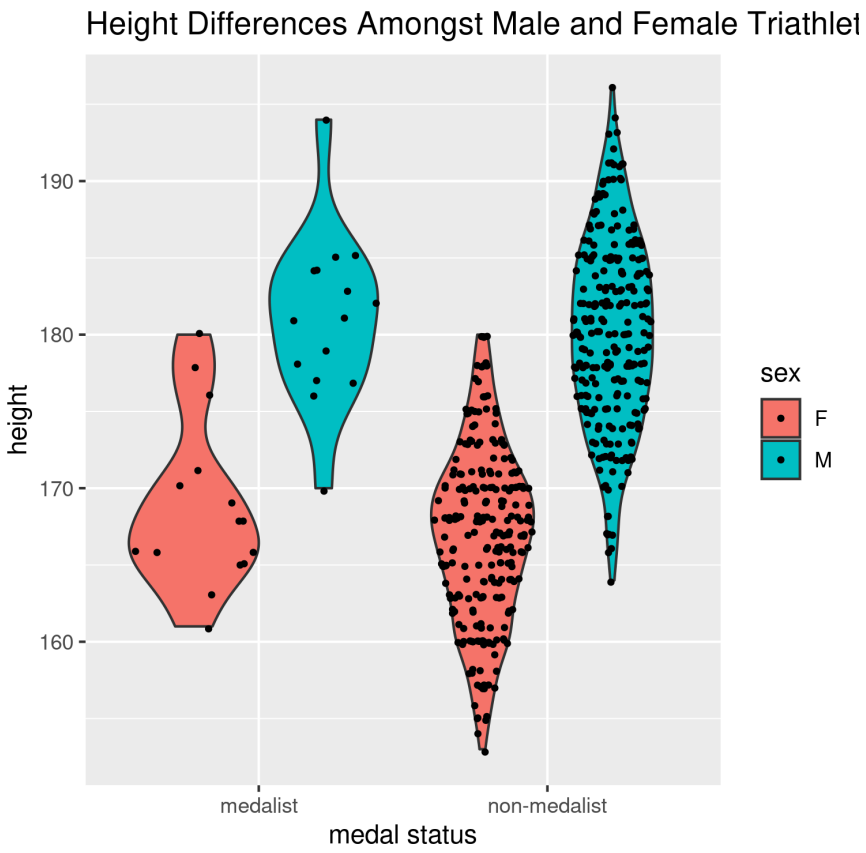
The United States and Great Britain have the greatest amount of female and male triatheltes competing in events Question 2: Are there height differences among triathletes between sexes or over time? To answer this question, we make a boxplot of height using year as a factor and fill by sex

```
# using x as factor to parse by year and labes to label the plot

ggplot(triathlon, aes(x = factor(year), y = height, fill = sex))+
  geom_boxplot() +
  labs(
    title = "Height Differences among Male and Female Triatheltes Over Time ",
    x = "year"
  )
```



Their is a clear difference in height between the sexes over time with a relationship seemingly being shown between height and year centered around a peak in 2008. Question 3: Are there height differences among triathletes that have medaled or not, again also considering athlete sex? To answer this question, we make a barplot of height using year as a factor and fill by sex

```
# fill of the plot is by sex and labs labels to label the plot.
library(ggforce)
ggplot(triathlon, aes(x = medalist, y = height, fill = sex))+
  geom_violin()+
  geom_sina(size = 0.75) +
  labs(
    title = "Height Differences Amongst Male and Female Triathletes based on
Medal Status ",
    x = "medal status"
  )
```



There does not seem to be a clear difference in height between the two respective sexes based on medal status Discussion: The top 3 countries that have the greatest amount of events competed for both sexes is The United States, Great Britain, and Switzerland. The countries that have the greatest differences in the number of events competed between the sexes is Japan, Ukraine, and Argentina. Also there appears to be countries which only has one sex competing in events such as Norway, South Korea, and Slovakia. It does not seems that there is an overall trend of numbers of events competed per sex that appears throughout the countries. Males and females triathletes show clear differences in height over time with their IQR never overlapping at any point in time. The year seems to have had an impact of the height of the male and female triathletes overall. With it showing a peak of height centered around 2008 as well as showing the most difference in height in 2008. We can see see from the tails of the boxplot that the year in which male and female triathletes have the least amount of differences also taking into consideration the IQR is the year 2012. The medal status of male and female triathletes does not seem to differ between medal status. We can see this in the violins in the fact that they have not moved significantly up or down or density hasn't chnaged significantly. While there are height differences between male and female triathletes overall in both status categories the status has no bearing on the height for each respective sex when comparing across the status categories. It does seem though that there are more triathletes that are in the

non-medalist category through the jitter in the plot. We can also see that there iws more variation and a more strong relationship in the non-medalist category. This is probably due to the fact that most triathletes do not win medals with this fact being constant.