Project 2

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')

olympic_gymnasts <- olympics %>%
  filter(!is.na(age)) %>%             # only keep athletes with known age
  filter(sport == "Gymnastics") %>%   # keep only gymnasts
  mutate(
    medalist = case_when(             # add column for success in medaling
      is.na(medal) ~ 'did not medal',   # NA values go to "did not medal"
      !is.na(medal) ~ 'medaled',          # non-NA values (Gold, Silver, Bronze) go to "medaled"
    ),
   sex = case_when(
     sex == 'F' ~ 'female',
     sex == 'M' ~ 'male',
    )
  )
```

More information about the dataset can be found at https://github.com/rfordatascience/tidytuesday/tree/master/data/2021/2021-07-27/readme.md (https://github.com/rfordatascience/tidytuesday/tree/master/data/2021/2021-07-27/readme.md) and https://www.sports-reference.com/olympics.html (https://www.sports-reference.com/olympics.html).

Question: Are there age differences for male and female Olympic gymnasts who were successful or not in earning a medal, and how has the age distribution changed over the years?

We recommend you use a violin plot for the first part of the question and faceted boxplots for the second question part of the question.

Hints:

- To make a series of boxplots over time, you will have add the following to your `aes()` statement: `group = year`.
- It can be a bit tricky to re-label facets generated with `facet_wrap()`. The trick is to add a `labeller` argument, for example:

```
+ facet_wrap(
   # your other arguments to facet_wrap() go here
   ...,
   # this replaces "TRUE" with "medaled" and "FALSE" with "did not medal"
   labeller = as_labeller(c(`TRUE` = "medaled", `FALSE` = "did not medal"))
 )
```

Introduction: The `olympics` dataset is a culmination of research performed by a group of Olympic enthusiasts who compiled data from both the Summer and Winter games, beginning with Athens 1896 all the way through Rio 2016. Each record within the dataset indicates the athlete who competed in an individual event. There are 15

columns, which indicate demographics about the individual athlete, such as their sex, age, height, weight, and team, in addition to the specific event the athlete competed in, which includes data points such as the City and Year where the event took place, the sport, the event, and the medal (if any) the athlete received. An additional boolean column, `medalist`, was created to indicate if the athlete received a medal or not.

In order to answer the question, I will use the following columns:

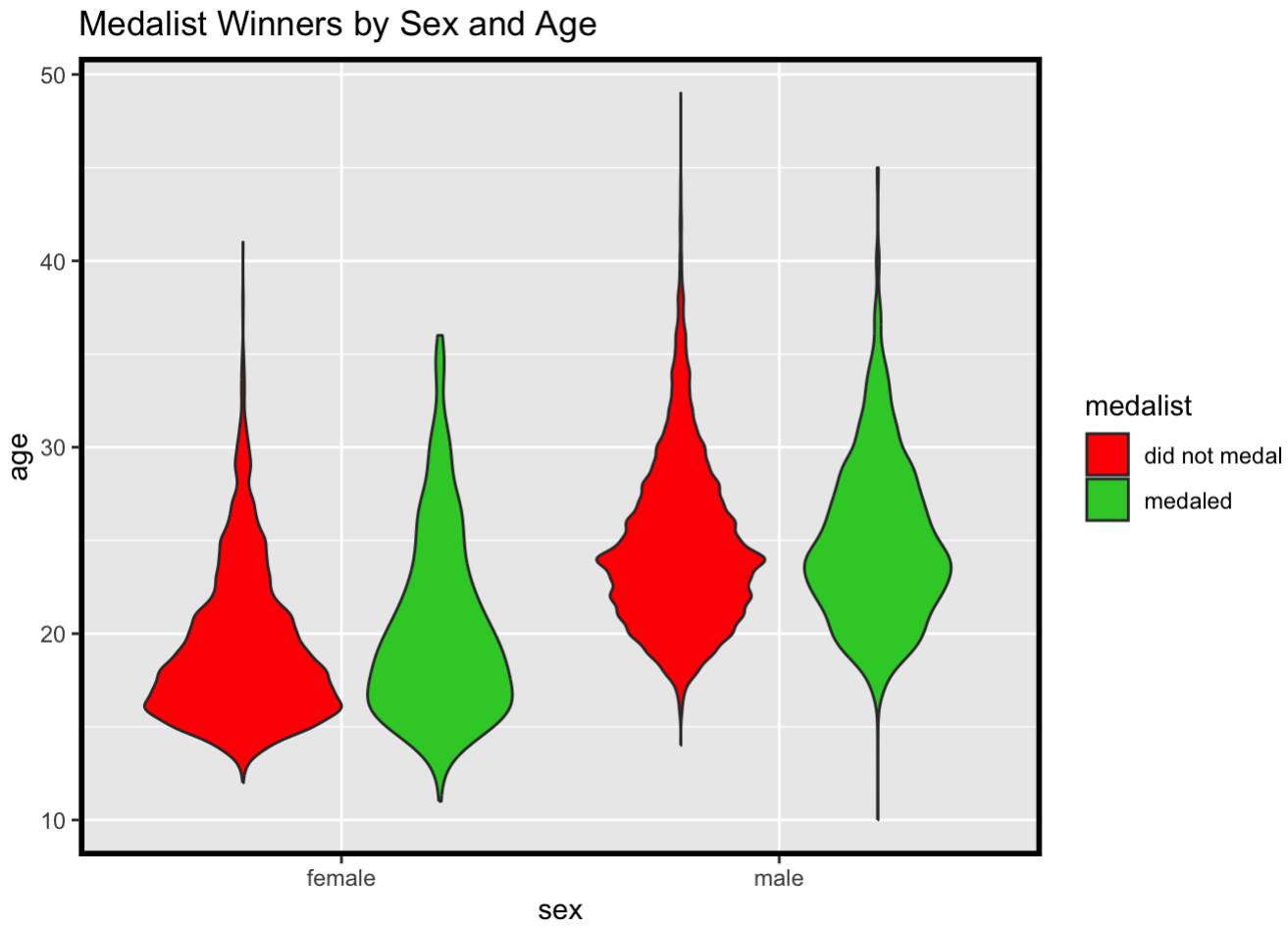
1. `age` : the athlete's age (integer)
2. `sex` : the athlete's sex (either "M" or "F")
3. `year` : the year of the Olympics (integer)
4. `medalist` : whether the athlete received a medal or not (*TRUE* if received either Gold, Silver, or Bronze, *FALSE* if otherwise)
5. `sport` : the sport the Olympian competed in (this is filtered in the dataset for "Gymnastics")

Approach: To visualize age differences for male and female Olympic gymnasts who were successful or not in earning a medal, I will use a violin plot to visualize the distribution of medalist winners by `sex` and `age`. Using a violin plot will best illustrate the age distribution of those who earned a medal versus those who did not. I am hypothesizing that all violin plots will show a "christmas tree" distribution for both medalist winners and those who did not win a medal, because it is likely that most olympian athletes are in the middle of their athletic career.

To visualize this same distribution but over years, it is best to use boxplots with `facet_grid` to show the four-dimension relationship, which will show the age distribution over the years for those who did not win a medal versus those who did, by `sex`. While a violin plot likely illustrates the distribution more clearly, leveraging violin plots here will be overwhelming for individuals viewing the graphic as there are many distributions with this granularity.

Analysis: To illustrate the age differences for male and female Olympic gymnasts who were successful or not in earning a medal, I am plotting the `age` distributions for each `sex` using a violin plot.

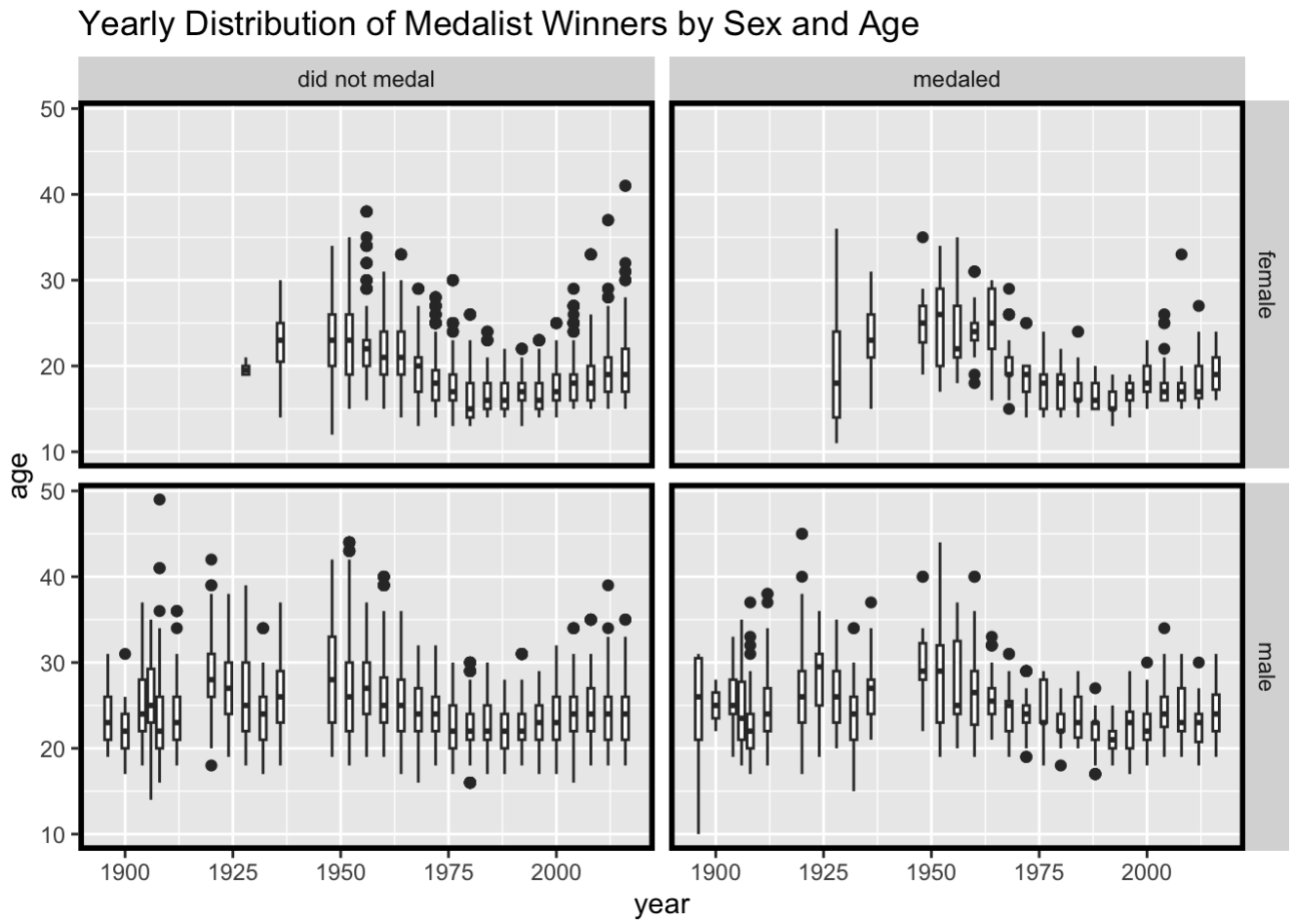
```
ggplot(olympic_gymnasts, aes(sex, age, fill=medalist)) +
    # Create violin
    geom_violin() +
    # Set Title and Axis Labels
    labs(title = "Medalist Winners by Sex and Age",
         x = 'sex',
         y = 'age') +
  #Specify Fill Color for each value of `medalist`
  scale_fill_manual(
    values = c("did not medal" = "#FF2400", "medaled"="#32CD32"),
  ) +
  #Create theme for final output
  theme(panel.border = element_rect(color = "black", fill=NA, linewidth=2))
```



To add the additional granularity of analyzing the differences for male and female Olympic gymnasts who were successful or not in earning a medal *over the years*, I am making a boxplot that shows the age distribution over the years and will facet by both `sex` and whether or not they earned a medal.

```
ggplot(olympic_gymnasts, aes(year, age, group=year)) +
    # Create Boxplot
    geom_boxplot() +
    # Facet by both `sex` and `medalist`
    facet_grid(sex ~ medalist) +

  #Create chart labels
  labs(title = "Yearly Distribution of Medalist Winners by Sex and Age",
         x = 'year',
         y = 'age') +
  #Create final output for theme
  theme(panel.border = element_rect(color = "black", fill=NA, linewidth=2))
```



Discussion: Upon examining the violin charts created, I was correct to hypothesize that the greatest distribution of gymnasts who have competed in the Olympic games are in the middle of their career. However, it is worth noting that the “middle” age for gymnasts differ between `sex`. Specifically, it appears the greatest distribution of female gymnasts are about 17 years old, while the greatest distribution of male gymnasts are about 24 years old, which may suggest females reach peak performance at an earlier age than males. It is also worth mentioning the range of the athletes’ ages between `sex`. It appears that males may have a longer “shelf life” when it comes to competing in the Olympics, noted by the oldest Olympian who did not win a medal, aged roughly 49 years old, and the youngest Olympian who won a medal, aged 10 years old, both who happen to be males.

Much of the information described above can also be gleaned from investigating the boxplot. With the added granularity of `year`, some observations become more apparent. Firstly, it appears that between the years of 1950 - early 1990, the typical age for male medalist winners declined, maybe suggesting a rise in athletic performance through genetics earlier on. However, there has been a slight uptick in the age lately. But perhaps the most striking observation from this visual is the “lack” of data for female athletes prior to 1925, which may suggest either that females were not allowed to perform in the olympics, or data for females who competed in the olympics were not tracked, which likely may due to equality during that time.