Project 2

This is the dataset you will be working with:

```
olympics <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesd
ay/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) ~ FALSE,           # NA values go to FALSE
      !is.na(medal) ~ TRUE            # non-NA values (Gold, Silver, Bronze) go to TRUE
    )
  )
```

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"))
  )
```

Statistics for introduction.

```
#number of athletes
length(unique(olympic_gymnasts[["id"]]))
```

```
## [1] 3665
```

```
#number of games
length(unique(olympic_gymnasts[["games"]]))
```

```
## [1] 29
```

```
#first game
x <- sort(olympic_gymnasts[["games"]])
x[1]
```

```
## [1] "1896 Summer"
```

```
#last game
x[length(x)]
```

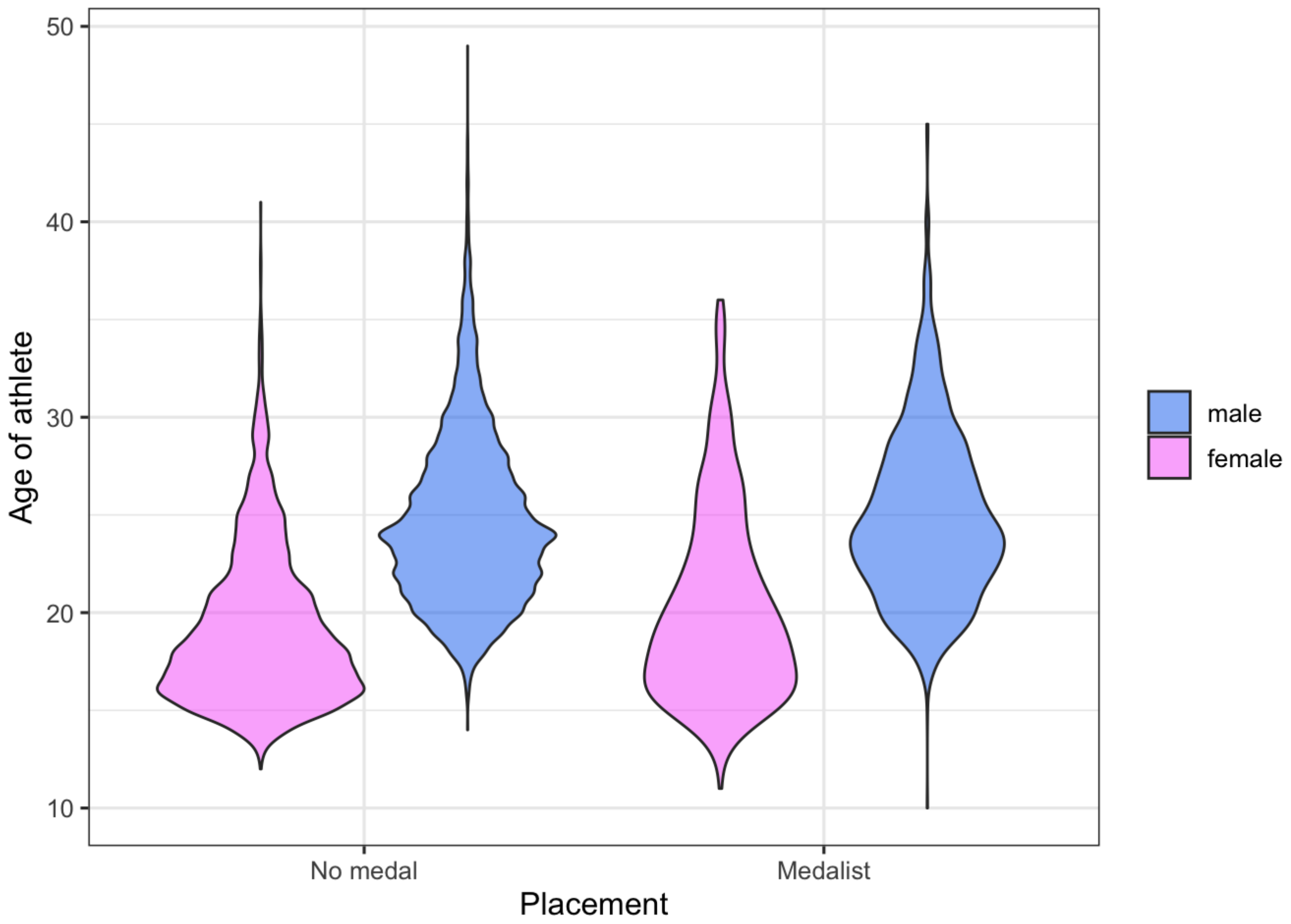
```
## [1] "2016 Summer"
```

Introduction: *The Olympic Games have been a sporting spectacle for centuries, dating back to the first ancient games held in during during the 8th century BC. Historians have poured over the event to document its evolution, and in the modern era it is time for us to do the same as Data Scientists. olympic_gymnasts is a dataset of the performance of 3665 gymnasts over 29 summer games from 1896 to 2016. To answer the following questions: “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 will use these specific columns: sex, age, year, medal, and medalist.*

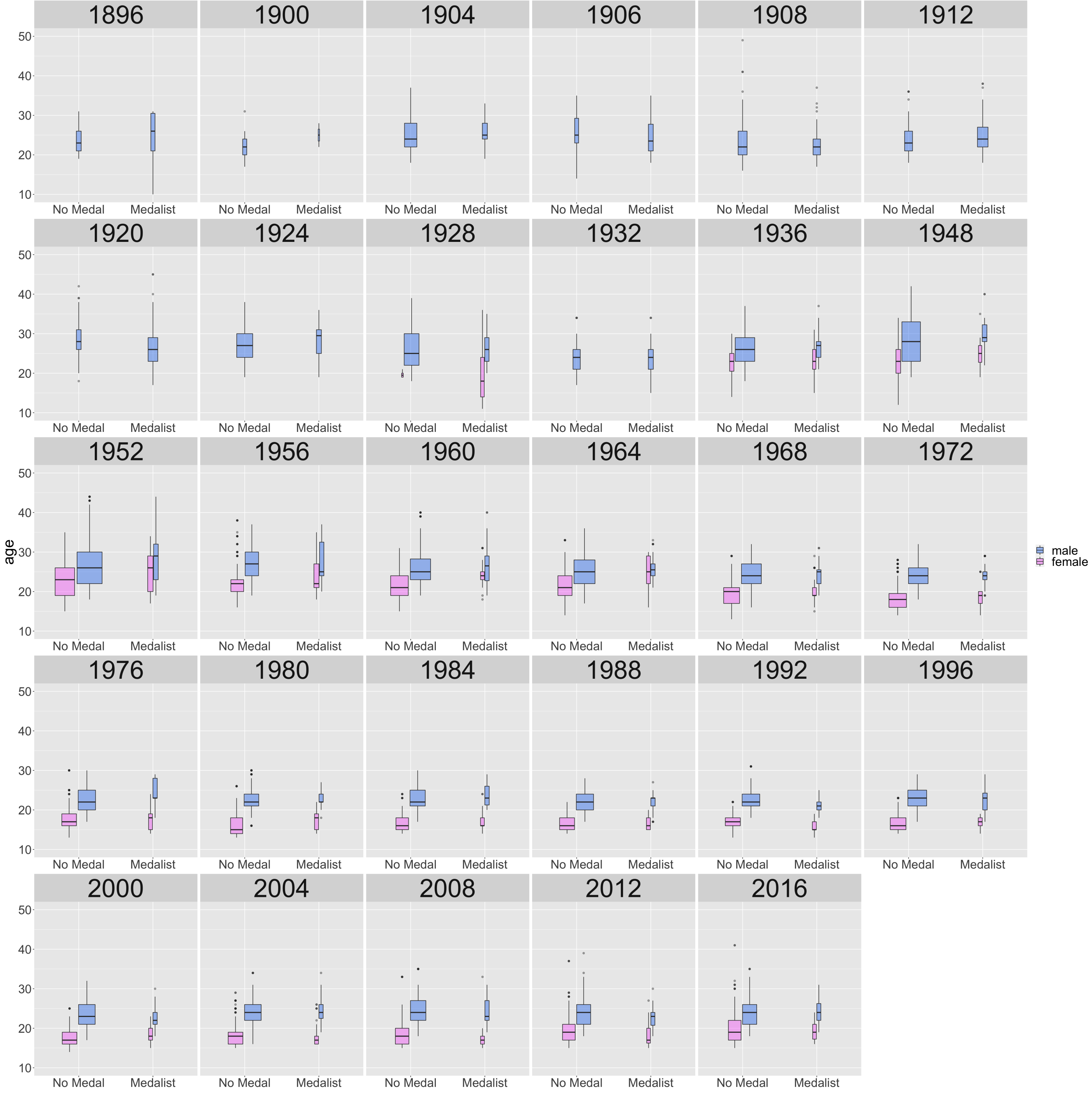
Approach: *To answer the first question, we’ll use a violin plot to visualize the age distribution of male and female medalists and non-medalists. The violin plot is used to show us a clearer view into the distribution, as opposed to a boxplot which requires more reading-into to understand. To answer the second question, we’ll use boxplots instead of violins (it might get a bit messy with the violins) and boxplots (with the parameter varwidth set to true) will allow us to make inferences based on the amount of values in boxplot. We’ll use facets to break down the boxplot into the individual years.*

Analysis:

```
ggplot(olympic_gymnasts, aes(factor(medalist), age, fill = factor(sex))) +
  geom_violin(alpha=.5) +
  scale_x_discrete(
    name = "Placement",
    labels = c("No medal", "Medalist")
  ) +
  scale_y_continuous(
    name = "Age of athlete"
  ) +
  scale_fill_manual(
    name = NULL,
    labels = c("male", "female"),
    values = c(`M` = "#0D71F0", `F` = "#F766F9")
  ) +
  theme_bw(12)
```



```
ggplot(olympic_gymnasts, aes(x = medalist, y = age, fill = sex))+
  geom_boxplot(
    alpha=.4,
    varwidth = TRUE
    ) +
  facet_wrap(
          .~year,
          scales = "free_x"
         ) +
  theme(
        strip.text.x = element_text(size = 50),
        text = element_text(size = 30)
        ) +
  coord_cartesian(
      ylim = c(10, 50)
    ) +
  scale_fill_manual(
    name = NULL,
    labels = c("male", "female"),
    values = c(`M` = "#0D71F0", `F` = "#F766F9")
    ) +
  scale_x_discrete(
    name= NULL,
    labels = c("No Medal","Medalist")
    )
```



Discussion: *Let's dig into the first plot. There seems to be no big difference between non-medalists and medalists in terms of age, as they are centered at around the same value. This could explained by the athletic nature of the events, requiring females and males to be in a specific age range at peak muscle mass development. Experience in the event could be a factor that causes this to go the other way (favoring older athletes) but that isn't the case in gymnastics. Gymnastic success depends individual ability rather than strategy. A good gymnast has to be able to balance, jump high, and soar through the air. There isn't a viable strategy other than "getting it right" as opposed to other Olympic events, such as team sports.*

While we're still on the first graph, let's talk about the differences between male and female athletes. It's clear that male athletes tend to be older, and female athletes tend to be younger. Why is this? It goes back to the age ranges of peak muscle mass between men and women. From "Essentials of Strength Training and Conditioning" a book *by the National Strength and Conditioning Association (NSCA) the following is said on the about the age gap on muscle development: “Peak muscle mass occurs between the ages of 16 and 20 years in females and between 18 and 25 years in males unless affected by resistance exercise, diet, or both.” This explains our differences in age.*

Moving onto the second plot, we have a lot to look at. First the obvious observations. The evolution of age, sex, and outcome in the Olympics changed drastically in 1928, as women were allowed to participate in the games, introducing female athletes. The amount of athletes that were awarded medals is much smaller than the non-medalists in almost every year (as seen by the width of the boxplots), as is the nature of the Olympic games where only the top three athletes are awarded medals. And the age gap talked about in the discussion of the first plot has seem to persisted across the years.

Let’s talk about some interesting observations. In 1928, there seems to be more medalists than non-medalists in women’s gymnastics. What’s up with that? According to the Wikipedia article for Gymnastics at the 1928 Olympic Games, “Only the team results (both combined and with respect to exercise) were published for the women, providing no information whatsoever about the capacities of the various individual women who competed here.” Interesting. And in 1932, there were no women gymnastic events, although there had been in the games prior. Another glance at Wikipeida shows: “There was apparently no women’s team gymnastics event like there was in the previous 1928 Olympics (which was the first Olympics where there was a women’s gymnastics competition) and like there would be for every single Summer Olympic games onward. No mention was ever made of this, nor was a rationale ever given anywhere in the Official Olympic report (or elsewhere), although there were women gymnasts who traveled to Los Angeles and participated in exhibition events at these games.”