Project 2

This is the dataset you will be working with:

```
olympics <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/maste
r/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?

Introduction: The dataset used in the study is the `olympics` dataset provided by *Sport Reference*; a company that compiles historical sporting data. The dataset contains data on olympic athletes from 1896 in Athens to 2016 in Rio.

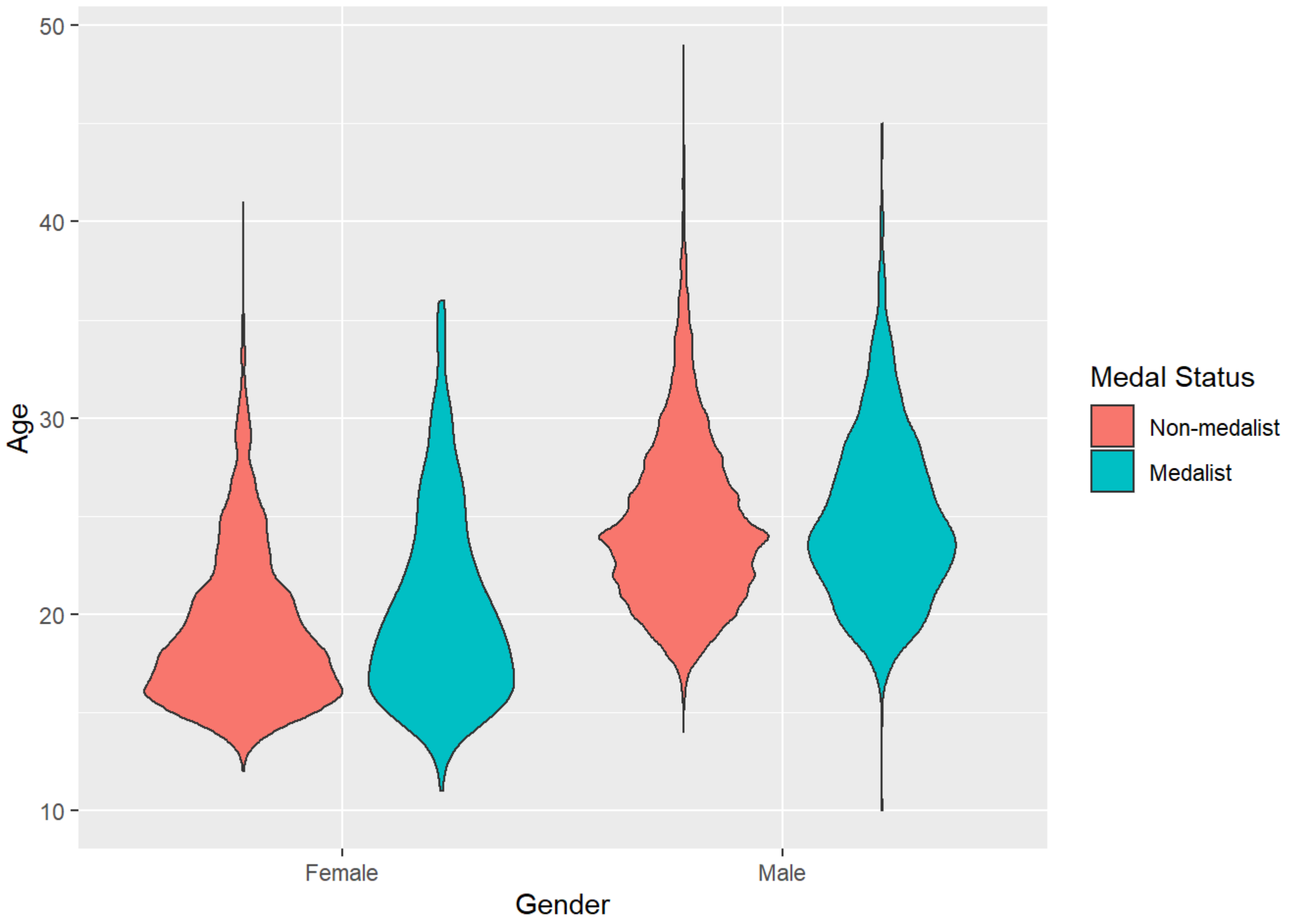
Within the dataset are various metrics on competing athletes. Examples are unique identifiers for each olympian (`id`), their name (`name`), their bodily traits (`height, weight`), what event they participated in (`event`), the year they participated (`year`), the city (`city`), and whether they achieved a medal or not (`medalist`). For this analysis, we will use the the `sex` , `age` , and `medalist` data columns

Approach: To start, the athletes were isolated by gender and whether or not they medaled. This helps to give context on the greater set of data by age without diving into the individual years yet. A simply violin plot shows much more granularity on the ages than does a more traditional boxplot, so that was the plot used.

For the following plot, to see greater details on the trends over time, facets across years was chosen. While grouping by year was a consideration, the resulting data was much too compressed to be easily interpretable. Violins were also considered, but for tracking the trend of medians over time, the extra noise did not make up for the benefit of more granularity. Therefore, boxplots were chosen for this plot.

Analysis: The first plot is the aforementioned violin plot of age vs gender and colored by whether they medaled or not.

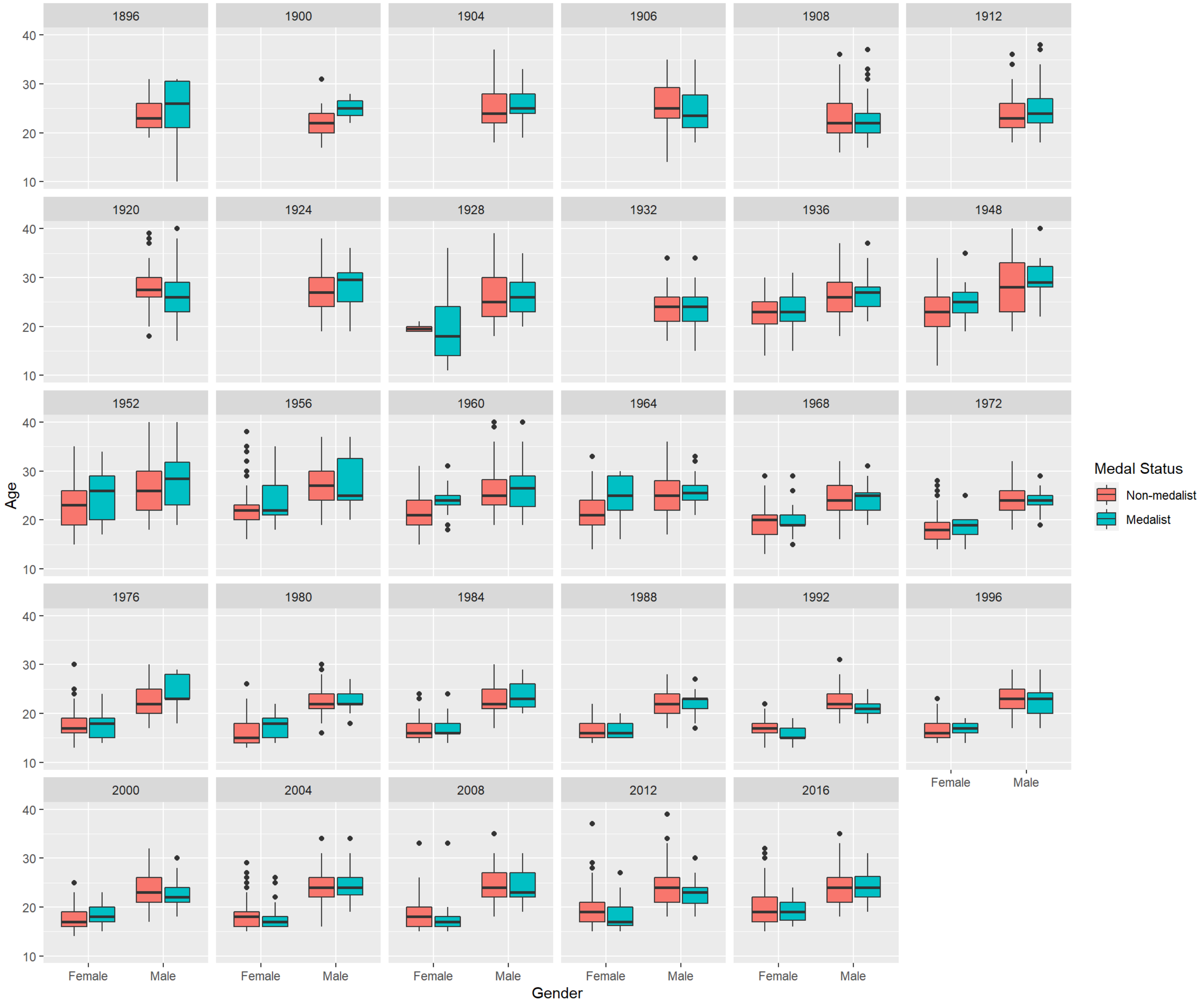
```
ggplot(olympic_gymnasts, aes(sex, age, fill = medalist)) +
  geom_violin() +
  scale_y_continuous(
    name = "Age",
  ) +
  scale_x_discrete(
    name = "Gender",
    labels = c("Female", "Male")
  ) +
  scale_fill_discrete(
    name = "Medal Status",
    labels = c("Non-medalist", "Medalist")
  )
```



The final plot facets the data by year to allow for tracking of age trends by gender and again whether they were a medalist or not.

```
ggplot(olympic_gymnasts, aes(sex, age, fill = medalist)) +
  geom_boxplot() +
  scale_y_continuous(
    name = "Age",
    limits = c(10, 40),
    breaks = c(10, 20, 30, 40)
  ) +
  scale_x_discrete(
    name = "Gender",
    labels = c("Female", "Male")
  ) +
  scale_fill_discrete(
    name = "Medal Status",
    labels = c("Non-medalist", "Medalist")
  ) +
  facet_wrap(
    vars(year)
  )
```

```
## Warning: Removed 47 rows containing non-finite values (stat_boxplot).
```



Discussion: By the first plot, the mean is roughly 8 years apart across gender; this is a strong difference when medal status is not accounted for. The violin plots for each gender across whether or not they medaled is quite similar; there is very little evidence that age has a factor in whether or not the olympian medals when isolating the data by gender. There appear to be more olympians for both men and women who continue to compete as the get older if they do not medal; it is likely that once a olympian medals, they feel more accomplished and are more willing to retire at a younger age.

By the second plot, we can first see that the olympian's age across gender holds true to the previous plot. Women consistently compete at younger ages than men do. Whether or not the athlete attains a medal or not again does not appear to be indicated or indicative of the athlete's age when isolated by gender.

The trend of the second plot can be very clearly seen by the medians as they track across the years. The scale of the y axis does not shift, so the ages can be clearly seen to drop over time. Athletes that compete are starting at younger ages over time.

Finally, it can also be seen more clearly by the boxplots that out of the athletes that do not medal, there are more outliers. It is likely that athletes that do not medal continue to try to attain a medal, or that older athletes become less capable of attaining them as they are older.