Project 2

This is the dataset you will be working with:

```
olympics <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/

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
    )
  )
```

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:

We are working with the `olympic_gymnasts` dataset, which was derived from the actual `Olympics` dataset by using appropriate filters. The filtered dataset contains 25528 gymnast records from all the Olympic games from 1896-2016. In this dataset, each row corresponds to one gymnast and there are 16 columns which provide Personal information, Host details and Sport particulars. Personal details include name, age, sex, height, weight, team. Host details include the place of host, season, year. Sport particulars include sports, events (Individual/Group) and the medal won (if any).

To answer the question, we will work with four variables, the age of the gymnast (column `age`), sex of the gymnast i.e, Male or Female (column `sex`), whether they won a medal or not, depicted by the mutated column `medalist` and the year in which the particular Olympic games was held (column `games`). The age of the gymnast is provided as a numeric value. The sex of the gymnast is encoded as M/F representing Male/Female. The games are provided as a year followed by the season in which it was held and the medalist is encoded as **True/False** where True means they won a Gold/Silver/Bronze medal in an event and False means they did not win any medal.

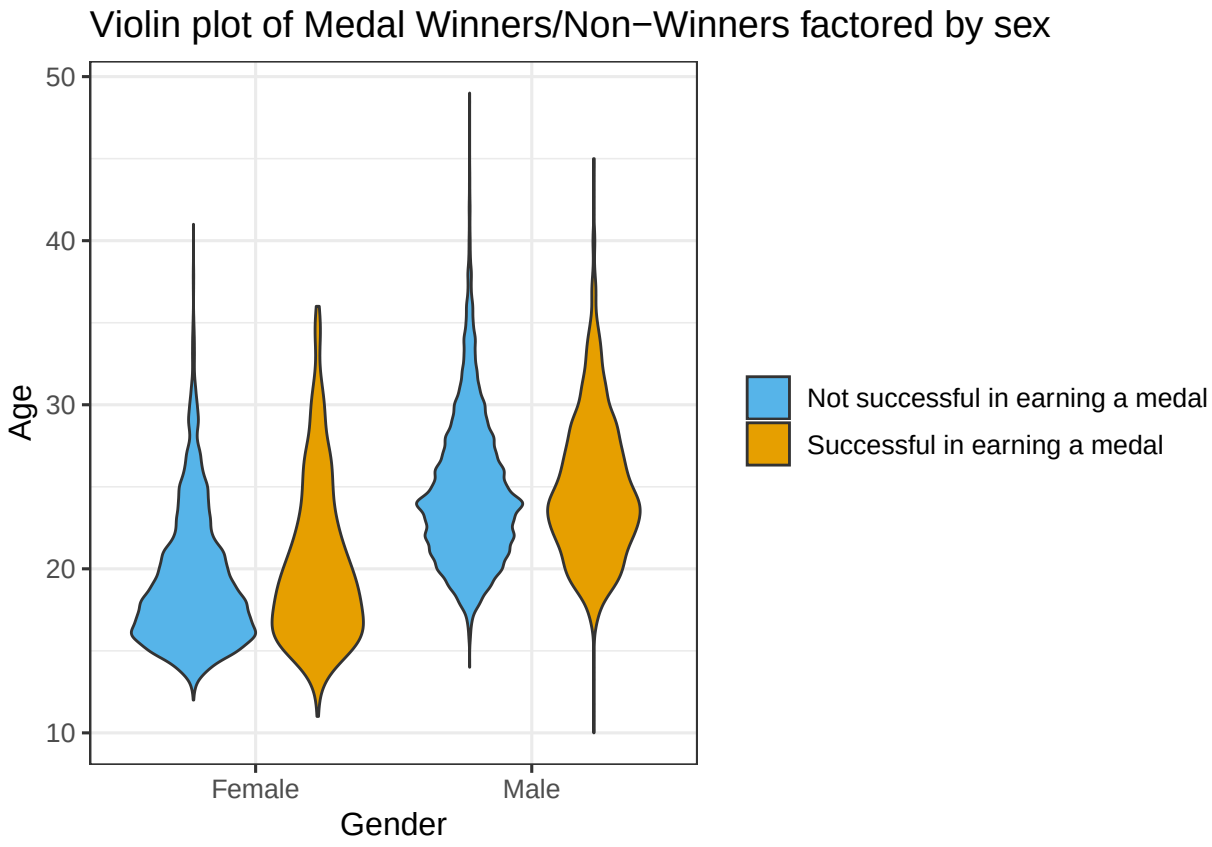
Approach:

Our approach is to show the distributions of age versus the two genders of the gymnasts i.e, male and female using violin plot (`geom_violin()`). We also separate out medal winners and those who didn't win, because medal winners are less in count compared to those who finish outside top 3. **Violins make it easy to compare multiple distributions side-by-side.**

One limitation of the violin plots is that they don't show us how many observations fall into each game of the Olympics. Therefore, we will visualize the number of medal winners/non-winners factored by the gender across each Olympic game with **faceted boxplots (`geom_boxplot()`).** Jointly, these two plots will allow us to answer the question.

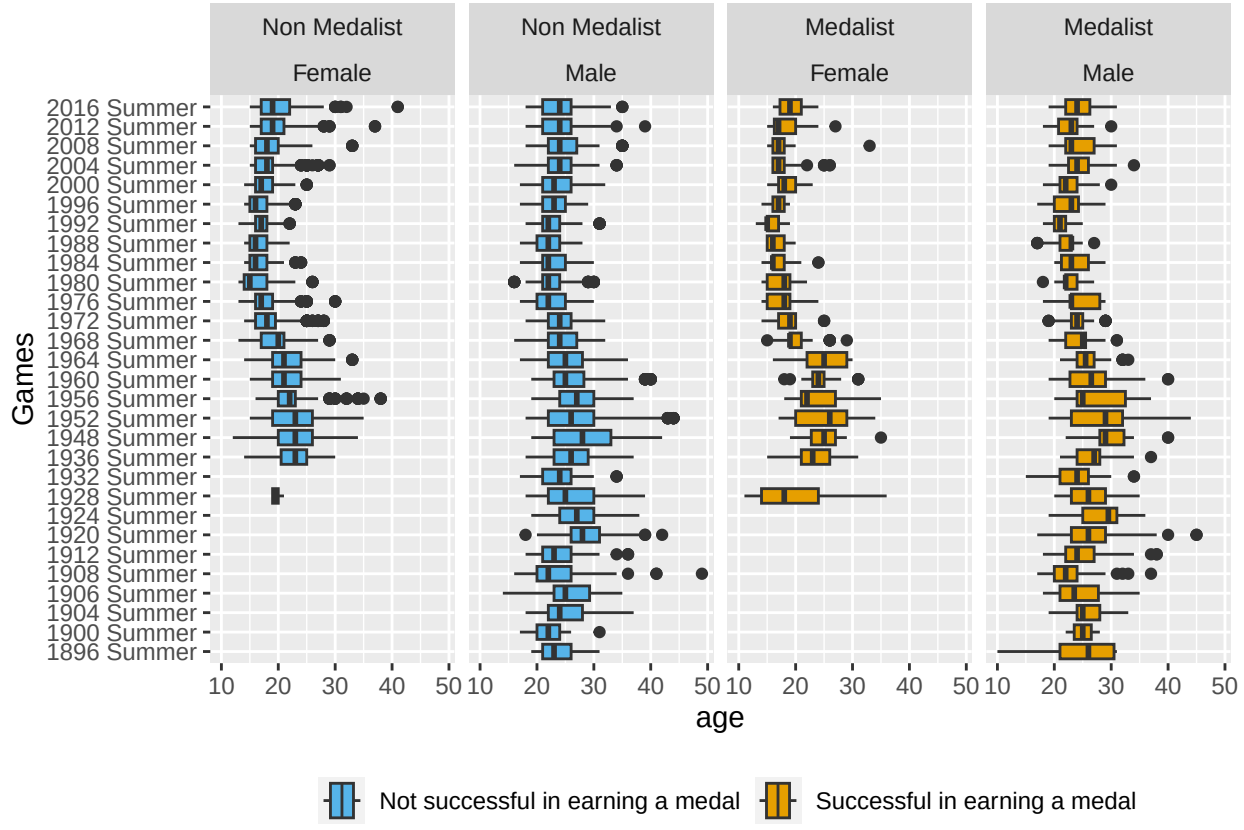
Analysis:
First we plot the medal winner vs non-winner distributions gender-wise as violins.

```
# We convert `sex` and `Medalist` into factors so R knows to treat them as discrete
# categorical variables
ggplot(olympic_gymnasts, aes(factor(sex), age, fill = factor(medalist))) +
  geom_violin() +
  scale_x_discrete(
    name = "Gender",
    # provide explicit labels so ggplot2 doesn't write 0 and 1
    labels = c("Female", "Male")
    ) +
  scale_y_continuous(name = "Age"
    ) +
  scale_fill_manual(
    name = NULL,
    # provide explicit labels as above for each category
    labels = c("Not successful in earning a medal", "Successful in earning a medal"),
    values = c(`TRUE` = "#E69F00", `FALSE` = "#56B4E9")
    ) +
  ggtitle("Violin plot of Medal Winners/Non-Winners factored by sex") +
  theme_bw(12)
```



Then we plot the Olympic gymnasts who were successful or not in earning a medal as a geom_boxplot factored by the gender in each Olympic game. We facet by both medalist, sex column grouped by year so we can see clearly the boxplot observations in each subset of the data.

```
# Again, need to convert variables into factors for ggplot2 to treat them
# as categorical.
ggplot(olympic_gymnasts,aes(x=age, y=games, fill=factor(medalist), group=year))+
  geom_boxplot() +
  scale_y_discrete(name = "Games"
    ) +
  scale_fill_manual(
    name = NULL,
    # provide explicit labels so ggplot2 doesn't write 0 and 1
    labels = c("Not successful in earning a medal", "Successful in earning a medal"),
    values = c(`TRUE` = "#E69F00", `FALSE` = "#56B4E9")
    ) +
  facet_wrap(
    vars(medalist, sex),
    nrow= 20,
    ncol = 6,
    # use `as_labeller` to convert 0 and 1 into meaningful labels
    labeller = as_labeller(c(`TRUE` = "Medalist", `FALSE` = "Non Medalist",
                             `M` = "Male", `F` = "Female"))
    ) +
  theme(legend.position = "bottom")
```



Discussion:

For those who were successful in earning a medal at the Olympics, the **female gymnasts were younger compared to the male gymnasts**. We can see this by comparing the orange violins in the first plot, where we see that they are shifted vertically relative to each other. It seems that some portion of the female medal winners were under 20 in age whereas the male medal winners were largely in the 20-30 age bracket. The same also holds true for the gymnasts who were not successful in earning a medal as well. The age difference is very clear by comparing the blue violins.

When we look at the breakdown of the age distribution of the gymnasts across all the Olympic games, we can clearly see that for female gymnasts, the age distribution of Medalists and Non-Medalists drops down from when they were allowed to participate. As for the male gymnasts, winners and non-winners, majority of the age distribution lies between 20-30 bracket and in the recent years the interquartile range has reduced.

Thus, the final answer to question is: Yes, there are age differences between male and female Olympic gymnasts who were successful or not in earning a medal. The age distribution has changed for the female gymnasts and their interquartile range has reduced significantly. For the male gymnasts, the age distribution has also reduced slightly.