Project 2

This is the dataset you will be working with:

```
olympics <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2

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
    )
  )
olympic_gymnasts
```

```
## # A tibble: 25,528 x 16
##       id name    sex     age height weight team   noc   games  year season city
##    <dbl> <chr>   <chr> <dbl>  <dbl>  <dbl> <chr>  <chr> <chr> <dbl> <chr>  <chr>
##  1    17 Paavo ~ M        28    175     64 Finla~ FIN   1948~  1948 Summer Lond~
##  2    17 Paavo ~ M        28    175     64 Finla~ FIN   1948~  1948 Summer Lond~
##  3    17 Paavo ~ M        28    175     64 Finla~ FIN   1948~  1948 Summer Lond~
##  4    17 Paavo ~ M        28    175     64 Finla~ FIN   1948~  1948 Summer Lond~
##  5    17 Paavo ~ M        28    175     64 Finla~ FIN   1948~  1948 Summer Lond~
##  6    17 Paavo ~ M        28    175     64 Finla~ FIN   1948~  1948 Summer Lond~
##  7    17 Paavo ~ M        28    175     64 Finla~ FIN   1948~  1948 Summer Lond~
##  8    17 Paavo ~ M        28    175     64 Finla~ FIN   1948~  1948 Summer Lond~
##  9    17 Paavo ~ M        32    175     64 Finla~ FIN   1952~  1952 Summer Hels~
## 10    17 Paavo ~ M        32    175     64 Finla~ FIN   1952~  1952 Summer Hels~
## # ... with 25,518 more rows, and 4 more variables: sport <chr>, event <chr>,
## #   medal <chr>, medalist <lgl>
```

More information about the dataset can be found at https://github.com/rfordatascience/tidytuesday/tree/master/data/2021/2021-07-27/readme.md and 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
    ...,
```

```r
    # this replaces "TRUE" with "medaled" and "FALSE" with "did not medal"
    labeller = as_labeller(c(`TRUE` = "medaled", `FALSE` = "did not medal"))
  )
```

Introduction:

To answer this question, we will plot the age-distribution by gender of the athlete and will also separate out the athletes who were successful in earning a medal or not. Along with it, will also plot the change in this age distribution over years, respectively.

We are working with the `Olympic Gynmnasts` dataset, which contains 25528 olympics gymnasts records from Athens 1896 to Rio 2016. In this dataset, each record corresponds to olympic gymnastics games played by atheletes and there are 16 columns providing information about the name, sex, age, height, weight of the athlete as well as events and games. It also includes information about the type of medals (Gold, Silver, Bronze, None) earned by the players. Information about the game include the name of the game, event-name, years and season during which it was conducted. It also includes the information about players whether they have earned medal(s) or not in the medalist column.

To answer the question, we will work with four variables, the age of the athlete (column `age`), whether the athlete has earned the award or not (column `medalist`), whether the athlete is a male or female (column `sex`) and in which years the athlete played (column `year`). The age column is provided as a numeric value, in numbers. The sex column is encoded as M/F, where M means male and F means female. Medalist column represents Whether the athlete has earned the award or not and is encoded as TRUE/FALSE, where True represents that athlete has won the medal, and False represents athlete has not won the medal. The values in year column ranges from 1896 to 2016 and it has numeric format.

Approach:

Our approach is to show the distributions of athletes age by their gender using violin plots (`geom_violin()`). We also separate out athletes who were successful in earning a medal or not, to understand whether there is an age-difference by gender in success or failure of earning a medal. Also, Violins make it easy to compare multiple distributions side-by-side.

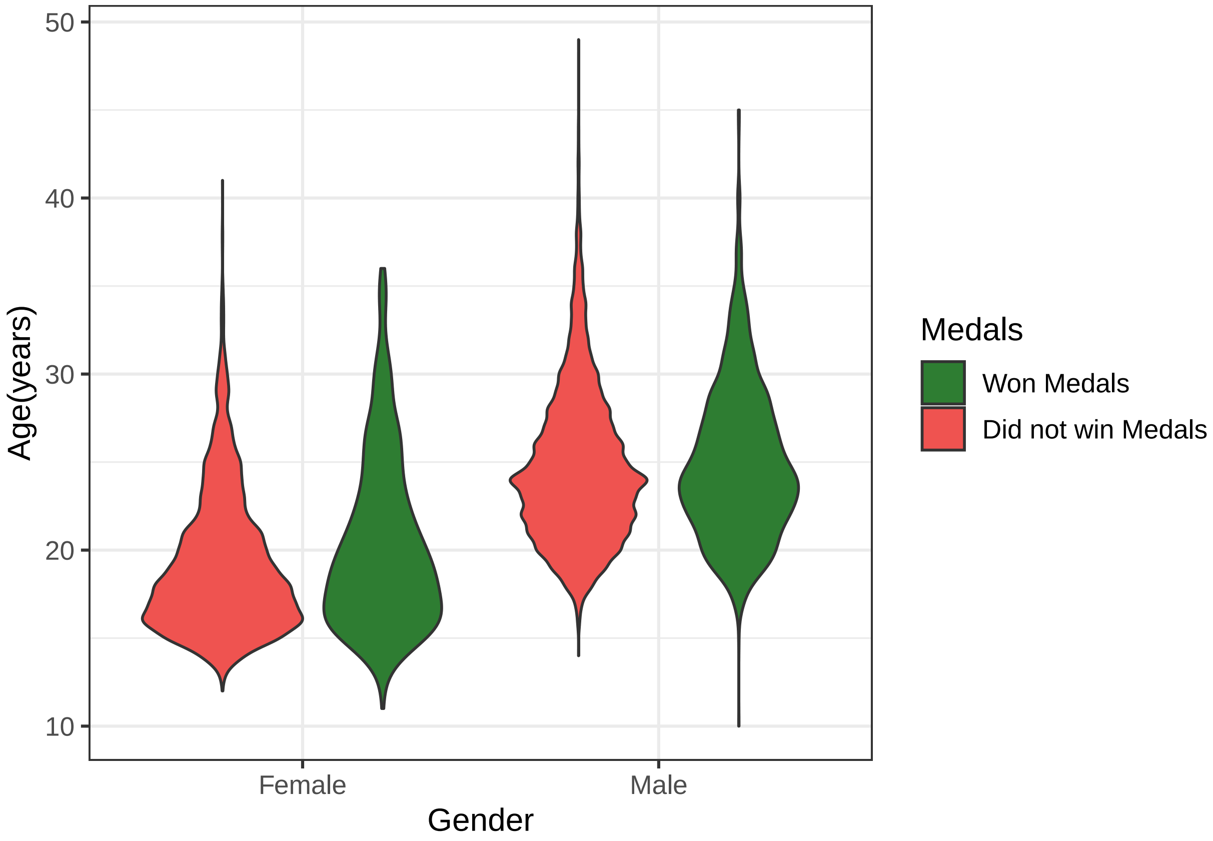
Further, To understand how this age-distribution (observed in violin plot) has changed over the span of years, we will plot this age distribution by years using boxplot.(`geom_boxplot`).

These two plots will allow us to answer the questions.

Analysis:

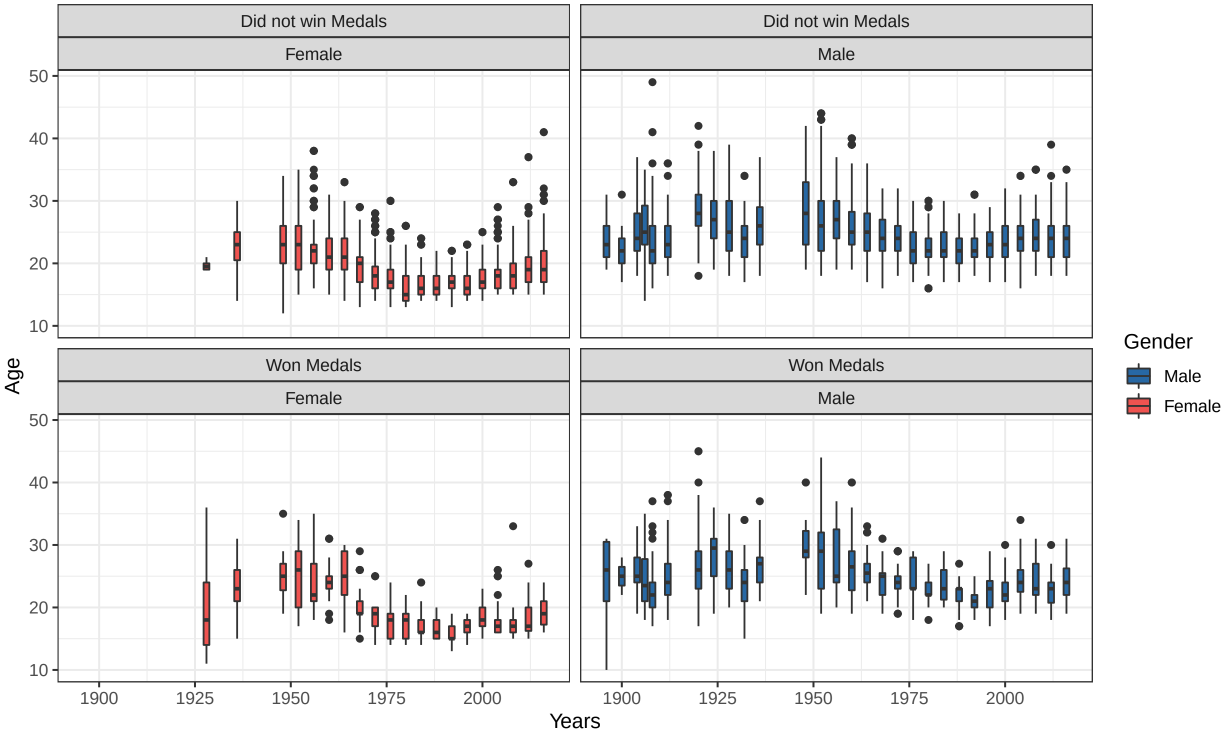
First we plot the age distributions as violins

```r
# Violin Plot
# We convert `sex` and `medalist` into factors so R knows to treat them as discrete categorical variabl

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 M and F
    labels = c(`M`="Male",`F`="Female")
  ) +
  scale_y_continuous(
    name = "Age(years)"
  ) +
  scale_fill_manual(
    name = "Medals",
    # provide explicit labels as below
    labels = c(`TRUE` = "Won Medals", `FALSE` = "Did not win Medals"),
    values = c(`TRUE` ="#2e7d32", `FALSE` = "#ef5350")
  ) +
  theme_bw(12)
```



Then we plot how the age-distribution (observed in the above plot) has changed over the years by using boxplot(geom_boxplot). We facet by medalist and sex of the athlete so we can clearly see how many observations there are in each subset of the data and also group them by years. We also separately account for sex of the athlete, to easily analyze how the age-difference has changed over the years by gender.

```
# BoxPlot
# Age distribution over the span of years.

ggplot(olympic_gymnasts, aes(year, age, group = year, fill=sex)) +
  geom_boxplot() +
  scale_x_continuous(
    name = "Years",
  ) +
  scale_y_continuous(
    name = "Age"
  ) +
  scale_fill_manual(
    name = "Gender",
    # provide explicit labels as below
    labels = c(`M`="Male", `F`="Female"),
    values = c(`M` = "#2768A4", `F` = "#ef5350")
  ) +
  facet_wrap(
    vars(medalist,sex),
    # use `as_labeller` to convert TRUE and FALSE into meaningful labels
    labeller = as_labeller(c(`TRUE` = "Won Medals", `FALSE` = "Did not win Medals",`M`="Male",`F`="Femal
  ) +
  theme_bw(12)
```



Discussion:

Age-difference between male and female olympic gymnasts is clearly visible by comparing the two violins plotted in female with the male. Average age of female gymnasts appears less than the average age of male gymnasts. However, whether gymnasts has won a medal or not, doesn't seem to have much effect on age-distribution. We can see this by comparing the two green violins with the two red violins, they are slightly shifted relative to each other but have otherwise a comparable shape. Age-distribution appears more affected by gender and little by whether a gymnast has earned a medal or not.

Also, Female gymnast have started competing after year 1925, whereas male gymnasts have started competing from 1896. Generally, Average age of male gymnasts seems greater than the average age of female gymnasts over the span of years as well. It can be compared by looking at the boxplots of color blue and pink. However, during 1950 to 1960, there is not a big variance in the average age of male and female gymnast. Moreover, Age-distribution continues to remain very little affected by whether a gymnast has earned a medal or not, except some outliers. It can be seen by comparing two blue boxplots, and two pink boxplots, respectively. Hence, age distribution appears to be affected largely by gender and varies slightly over the span of years.