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) ~ 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: Today, we will be looking at the Olympics data set, which contains 271,116 competitors over the entirety of the Summer and Winter Olympic Games (Athens 1896 to Rio 2016). We will be modifying the data set though, as we will be focusing solely on Olympic Gymnasts, which we see that there are 25,528 Olympic gymnasts in the data set. Each row an represents Olympic Gymnast that competed in a specific event (this means some Olympians are repeated in a given year as some compete in multiple events). For each athlete, the following is recorded: ID, name, sex, height, weight, Country, National Olympic Committee region, sport, event, medal results, Host, Season, and Year. We adjust the data set a little to make a new variable called “medalist”, which just takes the medal column and reads it as “True” if they had won any medal, or “False” if they did not win a medal. While using the new data set “Olympic Gymnasts”, we will be trying to answer the 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?” To answer the question, we will not need all the variables given to us. To answer the first part of the question, we will need the gymnast’s age of the year they competed (age), whether they are male or female (sex), and medal results (medalist). To answer the second part of the question, we will just look at the year the Olympics are held (Year), the age of the athlete (age), whether they are male or female (sex), and if they won a medal or not (medalist).

Approach: We will first make a table showing the number of observations of each sex and medal winners and non-winners just so we have an idea of how many are in each group. Our approach for the first part of the question, is to show the age distributions of Olympic medal winners and non winners of each sex using violin plots (geom_violin()). The violin plots allow us to see side-by-side comparisons of the age distributions of the groups. The second part of the question will be answered by making boxplots (geom_boxplots()) of age for each year for medal and non-medal winners of each sex. Using boxplots, we will be able to see the age trends of each group from the inception of gymnastics at the Olympics.

Analysis: First, a look at what the data set looks like and the counts of medal winners and non-medal winners of each sex.

```
olympic_gymnasts %>%
  select(name,age, sex,medalist, year)
```

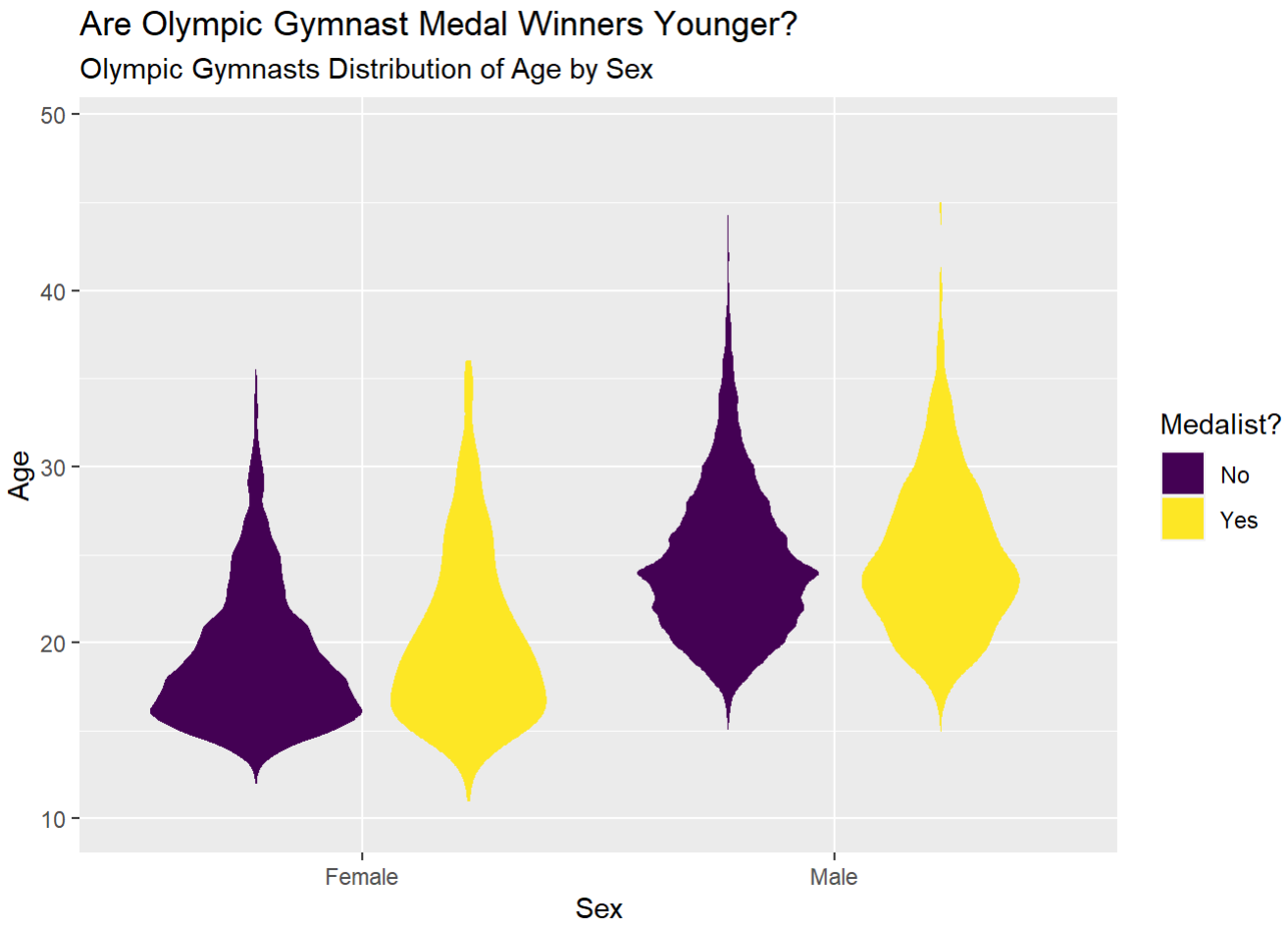
```
## # A tibble: 25,528 x 5
##    name                      age sex   medalist  year
##    <chr>                   <dbl> <chr> <lgl>    <dbl>
##  1 Paavo Johannes Aaltonen    28 M     TRUE      1948
##  2 Paavo Johannes Aaltonen    28 M     TRUE      1948
##  3 Paavo Johannes Aaltonen    28 M     FALSE     1948
##  4 Paavo Johannes Aaltonen    28 M     TRUE      1948
##  5 Paavo Johannes Aaltonen    28 M     FALSE     1948
##  6 Paavo Johannes Aaltonen    28 M     FALSE     1948
##  7 Paavo Johannes Aaltonen    28 M     FALSE     1948
##  8 Paavo Johannes Aaltonen    28 M     TRUE      1948
##  9 Paavo Johannes Aaltonen    32 M     FALSE     1952
## 10 Paavo Johannes Aaltonen    32 M     TRUE      1952
## # ... with 25,518 more rows
```

```
olympic_gymnasts %>%
  count(sex, medalist)
```

```
## # A tibble: 4 x 3
##   sex   medalist     n
##   <chr> <lgl>    <int>
## 1 F     FALSE     8349
## 2 F     TRUE       695
## 3 M     FALSE    14992
## 4 M     TRUE      1492
```

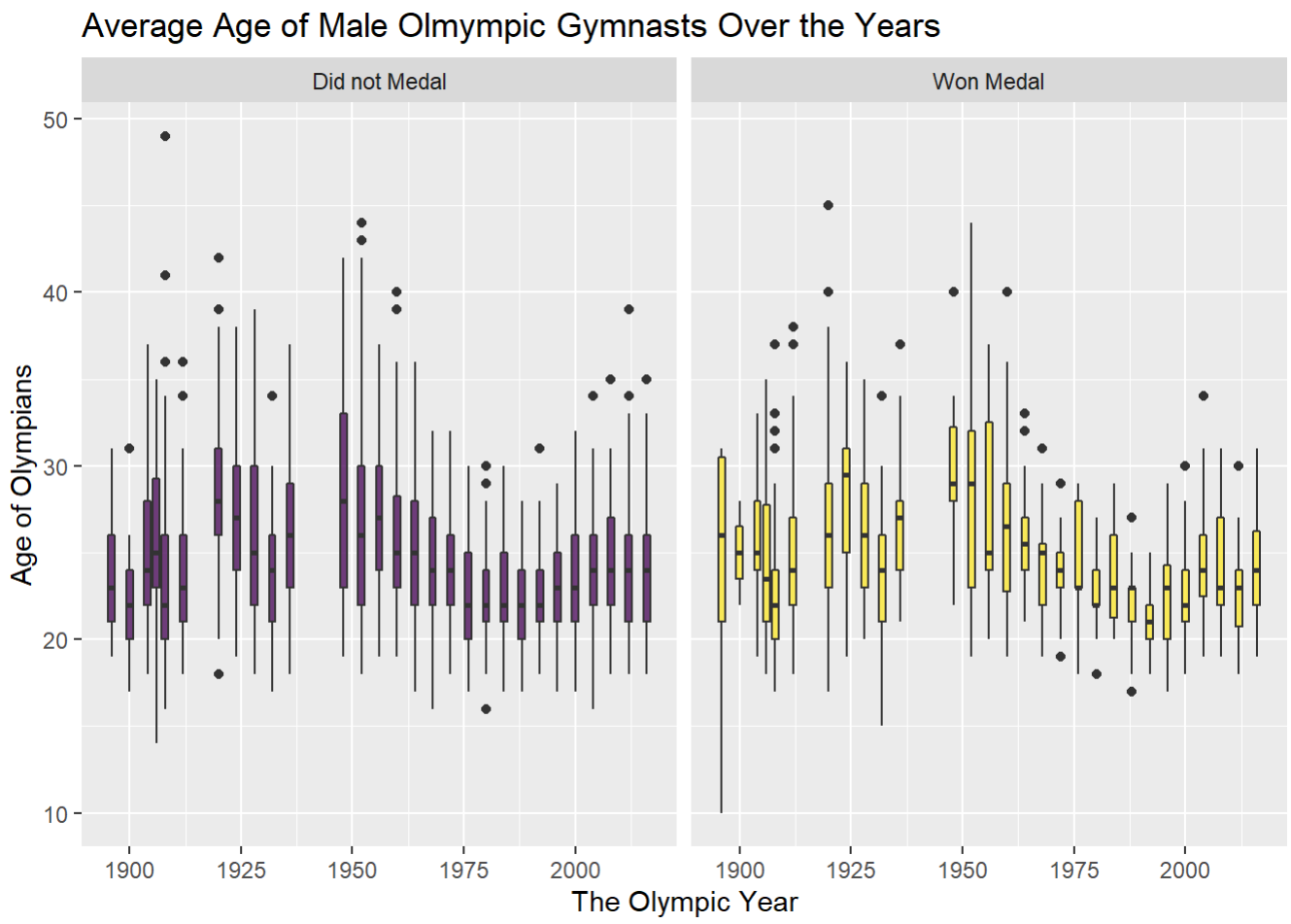
Then, the violin plots showing the age dsitributions among each group.

```
ggplot(olympic_gymnasts, aes(sex, age, fill = medalist)) +
  geom_violin(color = NA, trim = TRUE) +
  labs(title = "Are Olympic Gymnast Medal Winners Younger?", subtitle = "Olympic Gymnasts Distri
bution of Age by Sex") +
  scale_x_discrete(name = "Sex", labels = c("Female", "Male")) +
  scale_y_continuous(name = "Age") +
  scale_fill_viridis_d(name = "Medalist?", labels = c("No","Yes"))+
  theme_grey()
```

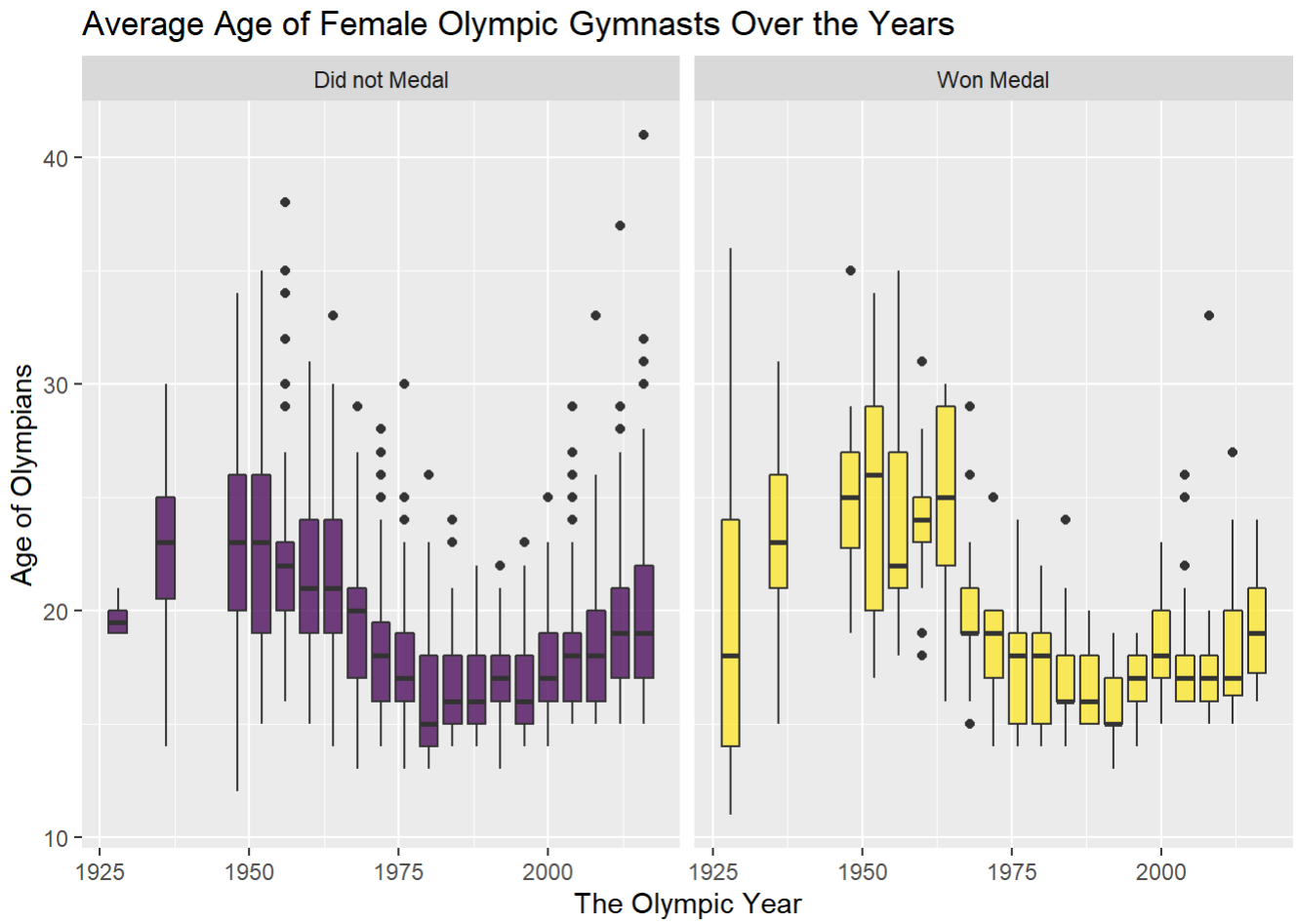


And lastly, four faceted boxplots showing the age trends of medal and non-medal winners of each sex.

```
olympic_gymnasts %>%
  filter(sex == "M") %>%
  ggplot(aes(year, age, group = year, fill = medalist)) +
   geom_boxplot() +
    facet_wrap(vars(medalist), labeller = as_labeller(c(`TRUE` = "Won Medal", `FALSE` = "Did not
Medal"))) +
  scale_fill_viridis_d(alpha = 0.75, guide = "none") +
  labs(title = "Average Age of Male Olmympic Gymnasts Over the Years", x = "The Olympic Year", y
= "Age of Olympians")+
  theme_grey()
```




```
olympic_gymnasts %>%
  filter(sex == "F") %>%
  ggplot(aes(year, age, group = year, fill = medalist)) +
    geom_boxplot() +
    facet_wrap(vars(medalist), labeller = as_labeller(c(`TRUE` = "Won Medal", `FALSE` = "Did not
Medal"))) +
  scale_fill_viridis_d(alpha = 0.75, guide = "none") +
  labs(title = "Average Age of Female Olympic Gymnasts Over the Years", x = "The Olympic Year",
 y = "Age of Olympians")+
  theme_grey()
```



Discussion: The violin plots help us answer the question, “Are there age differences for male and female Olympic gymnasts who were successful or not in earning a medal?”. Looking at both groups (male and female), we can see that medal and non-medal winners share the same age distribution. For the female group, we see both violins start off thin and get more dense around 17 years old and then slowly get thinner as the years go on until they stop at around 35 years old. For the male group, we see both violins start around 15 and slowly get more dense up to the age of 24, and then slowly decrease to the age of 40. We see that for both sexes, medal and non-medal winners share the same age distribution trend. We should run a statistical test though to compare age amongst groups.
The two boxplot visuals allow us to answer the question, “how has the age distribution changed over the years?”. While analyzing the boxplots of each sex, we see a couple of trends of the medal and non-medal winner age distributions. When gymnastics was first implemented into the Olympics for each sex, we see a low of about 25 years old for males (early 1900s) and 18 years old for females (1920s). We see the age of all groups increase increase up to the 1950s (29 for males and 25 for females) and then decline to the 1980s (22 for males and 15 for females) and then slowly increase and even out to present day (24 for males and 18 for females). We see the trends for each group between sex is the same and that typically the younger group is medal winners.
Thus to answer both questions, between the sexes, I do not think there is an average age difference amongst medal and non-medal winners. And for the age distribution, all 4 groups shared a similar trend of being young when the Olympics first introduced gymnastics and gradually getting older up to the 1950s, and then finally declining and evening out after the 1980s. I think we see similar trends between the groups because of the types of athletes competing in those Olympic years.