Project 1

Ethan Chang - ehc586

This is the dataset you will be working with:

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

olympics_alpine <- olympics %>%
  filter(!is.na(weight)) %>%              # only keep athletes with known weight
  filter(sport == "Alpine Skiing") %>%    # keep only alpine skiers
  mutate(
    medalist = case_when(                 # add column to
      is.na(medal) ~ FALSE,               # NA values go to FALSE
      !is.na(medal) ~ TRUE                # non-NA values (Gold, Silver, Bronze) go to TRUE
    )
  )
```

`olympics_alpine` is a subset of `olympics` and contains only the data for alpine skiers. More information about the original `olympics` 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.

For this project, use `olympics_alpine` to answer the following questions about the weights of alpine skiers:

1. Are there weight differences for male and female Olympic skiers who were successful or not in earning a medal?
2. Are there weight differences for skiers who competed in different alpine skiing events?
3. How has the weight distribution of alpine skiers changed over the years?

You should make one plot per question.

Introduction: For this project, we work with the Olympic Games dataset, a dataset consisting of information for Olympic competitors from Athens 1896 to Rio 2016. Each row in the dataset corresponds to a participant from a specific event with multiple columns denoting that participant's physical characteristics, associations, events, and results. More specifically, the dataset contains a participant's id, name, sex, age, height, weight, team, nationality, the games, year, season, city, sport, and event they played, and type of medal they won, if any.

To answer the three questions provided above, we look at a specific subset of the `olympics` dataset, `olympics_alpine`, as we're only interested in the weight distribution of alpine skiers, those who have participated in the sport Alpine Skiing. From this dataset, we work with 5 variables, the participant's `weight` and `sex`, whether or not they medaled (`medalist`), the Alpine Skiing events (`event`), and `year` of the event they participated in. The weight is provided in kilograms, while the sex is given as either an M (male) or F (female). Medalists are presented as a boolean, with TRUE representing medalists and FALSE representing non-medalists. The events are listed as a character containing the name of the event, and the year is listed as a numeric value.

Approach: To show the distribution of weights between male and female skiers who medaled and did not medal, we used violin plots (`geom_violin()`) as it makes it easy to compare multiple distributions. We separated the weight distributions by `sex` and faceted it by `medalist` to allow for comparisons between these variables to answer the question.

To show the weight distributions between skiers who participated in different alpine skiing events, we used

boxplots (`geom_boxplot()`) as it allows us to directly compare distributions and quartiles between different events easily. We would color map it by sex as men and women have different weights on average, allowing for better comparisons between events.

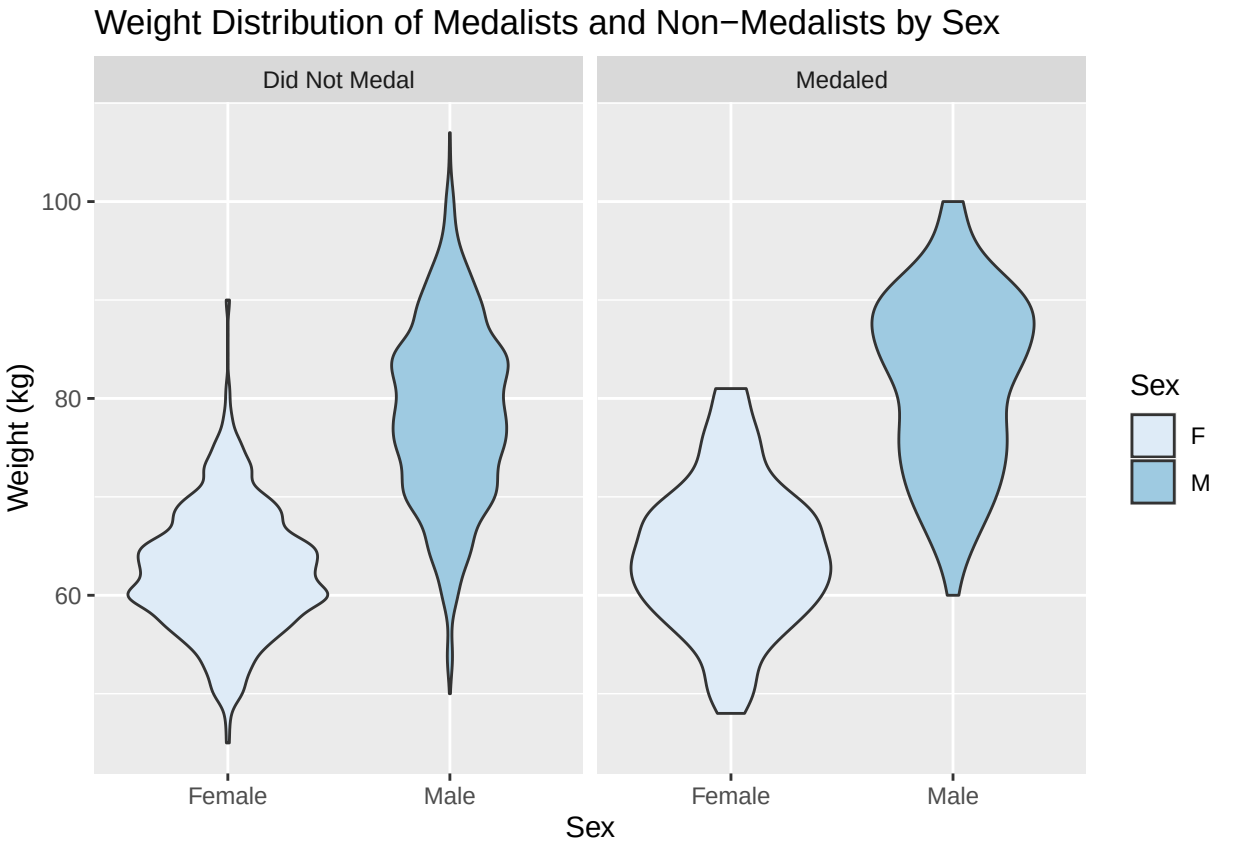
Finally, to show the weight distribution of alpine skiers over time, we used boxplots again. This allowed us to show multiple weight distributions and quartiles over multiple years from earliest to latest to facilitate comparisons.

Analysis:

Question 1: Are there weight differences for male and female Olympic skiers who were successful or not in earning a medal?

To answer this question, we plot the weight distribution as violins, separated by both sex and whether they medaled or not.

```
ggplot(olympics_alpine, aes(sex, weight, fill = sex)) +
  geom_violin() +
  # separates plots between medalists and non-medalists
  facet_wrap(~medalist,
             labeller = as_labeller(c(`TRUE` = "Medaled", `FALSE` = "Did Not Medal"))) +
  labs(x = "Sex", y = "Weight (kg)", fill = "Sex",
       title = "Weight Distribution of Medalists and Non-Medalists by Sex") +
  scale_x_discrete(labels = c("Female", "Male")) +
  scale_fill_brewer()
```

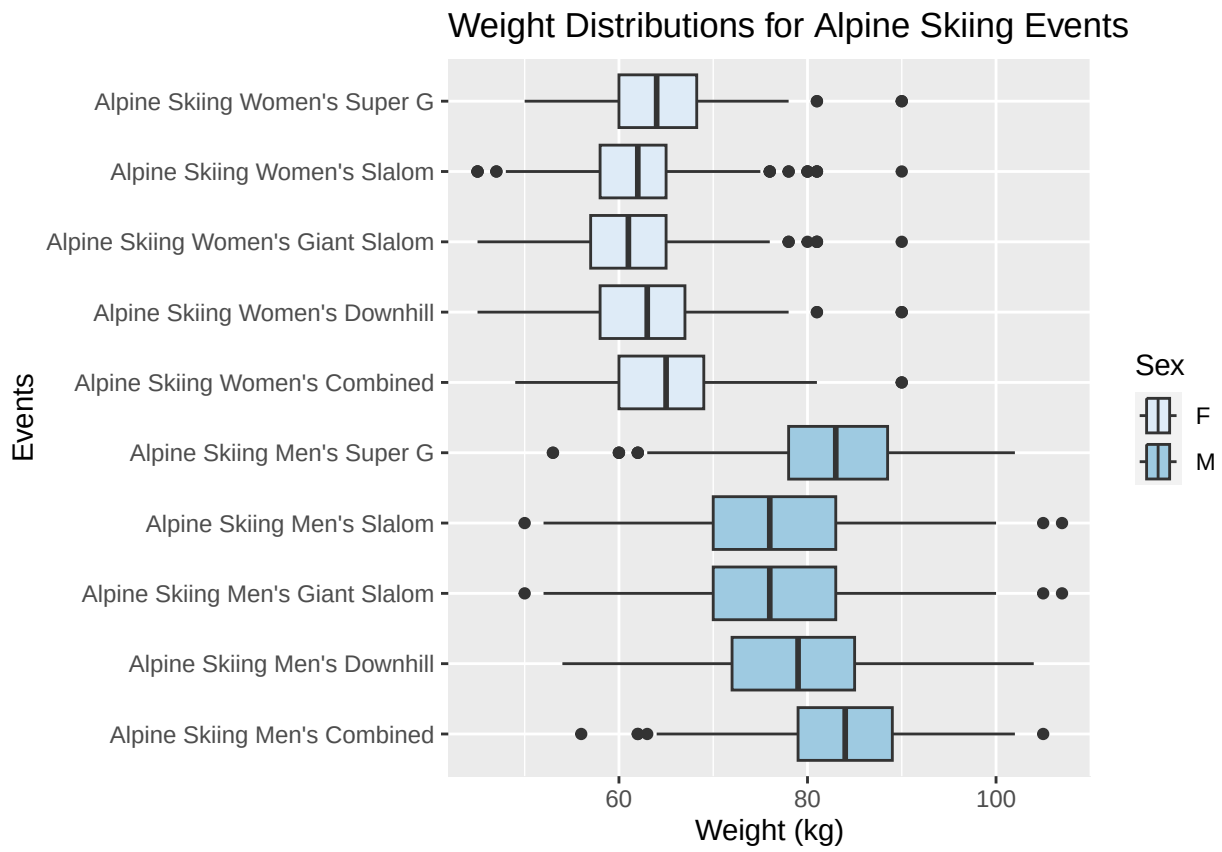


There is a clear difference between male and female weights, regardless of whether they medaled or not, with medaled weights appearing to be slightly higher.

Question 2: Are there weight differences for skiers who competed in different alpine skiing events?

To answer this question, we made multiple boxplots of the weight distribution per event, color mapped by sex.

```
ggplot(olympics_alpine, aes(weight, event, fill = sex)) +
  geom_boxplot() +
  labs(x = "Weight (kg)", y = "Events", fill = "Sex",
       title = "Weight Distributions for Alpine Skiing Events") +
  scale_fill_brewer()
```

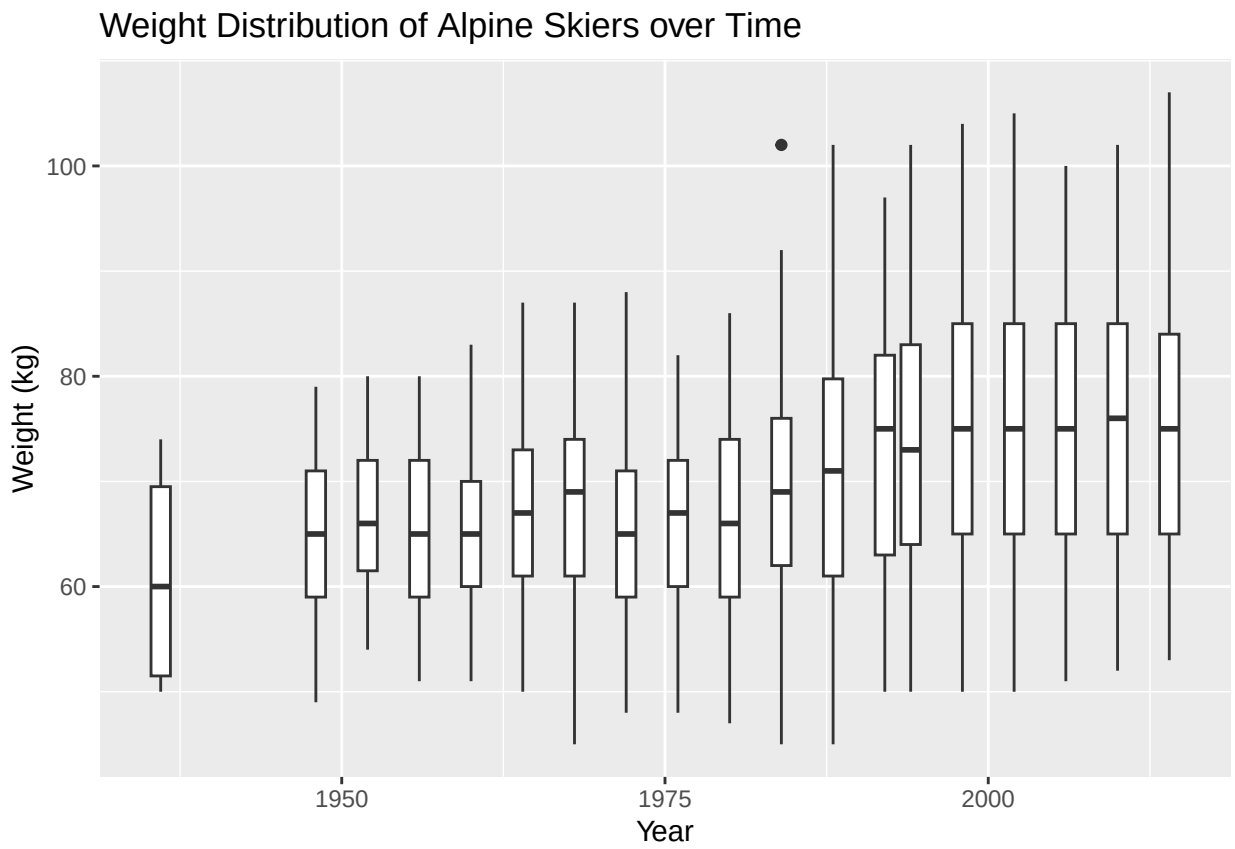


Again, there is a clear difference between male and female weight distributions with male distributions being higher on average and events maintaining similar distributions overall relative to each sex.

Question 3: How has the weight distribution of alpine skiers changed over the years?

To answer this question we simply showed multiple weight distributions over time in the form of boxplots.

```
ggplot(olympics_alpine, aes(year, weight)) +
  geom_boxplot(aes(group = year)) +
  labs(x = "Year", y = "Weight (kg)",
       title = "Weight Distribution of Alpine Skiers over Time")
```



The overall weights of skiers appear to increase over time as the medians and quartiles appear to increase.

Discussion: Based on the first figure produced in our analysis, it can be seen that for both medalists and non-medalists, there is a clear difference in weight between males and females. As expected, males tend to have higher weights (around 80 kg) than females (around 60 kg) which we can see in the violin plots. When comparing the weights of medalists vs non-medalists for both sexes, overall weight distributions seem to remain the same, but there are a few minor unique differences that can be seen. Medalists tend to have a more uniform distribution of weights with a slightly higher average weight, whereas non-medalists have a more sporadic, multimodal distribution of weights with a slightly lower average weight. These trends could be a result of there being more non-medalists than medalists, resulting in a more uneven distribution of weights for non-medalists, and a higher concentration of similar weights for medalists, though further analysis or statistical tests would have to be done to check this.

From our second figure, when separating the events by male and female, the same trend can be seen where all female events have lower weight distributions than male events as the boxplots for all female events have upper quartiles that are at most equivalent to the lower quartiles of all male events. When comparing the weight distribution of each event per sex, it can be seen that each event has a similar distribution relative to other events for its own sex. For example, both Slalom and Giant Slalom events have similar weight distributions to each other and are the lowest weight distribution overall for both male and female. Similarly, both Super G and Combined events have similar weight distributions and have the highest weight distribution overall for both sexes, with Downhill trailing slightly behind them. This implies that there are differences in weight between different skiing events, and that different events may attract more people of different weights.

For our last figure, when viewing the distribution of weights across the years, it can be seen that there has been a general increase in the overall weight of alpine skiers with its distribution fluctuating a bit. This is shown in the boxplots as the medians tend to create an upward trend over time and have relatively centered, unimodal distributions, with ranges and medians fluctuating throughout. This pattern of weight distribution changes could be a result of newer participants with different weights, either from different training regimes,

diets, genetics, or even different sex distributions. There are lots of different variables that could potentially explain this change in weight distribution over time, and we would have to conduct further analysis to determine which variables were most relevant in influencing these results.