Project 1

Wenxuan Zhou wz4388

This is the dataset you will be working with:

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

members_everest <- members %>%
  filter(peak_name == "Everest") %>% # only keep expeditions to Everest
  filter(!is.na(age)) %>%     # only keep expedition members with known age
  filter(year >= 1960)        # only keep expeditions since 1960
```

More information about the dataset can be found at https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-09-22/readme.md and https://www.himalayandatabase.com/.

Part 1

Question: Are there age differences for expedition members who were successful or not in climbing Mt. Everest with or without oxygen, 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
    ...,
    # this replaces "TRUE" with "summited" and "FALSE" with "did not summit"
    labeller = as_labeller(c(`TRUE` = "summited", `FALSE` = "did not summit"))
  )
```

Introduction: *We are working with the* `members_everest` *dataset, which contains records specifically for expeditions to Mount Everest since 1960 and expedition members of known age. In this dataset, each row corresponds to one individual who participated in the expedition, and there are 21 columns providing information about the individual and the expedition. Information about the individual includes member id, sex, age, citizenship, expedition role, whether the person was hired by the expedition, elevation highpoint of the person, whether the person was successful in climbing to a peak, and whether the person used oxygen. Information about the expedition includes the expedition id, peak id, peak name, year of expedition, and season of expedition.*

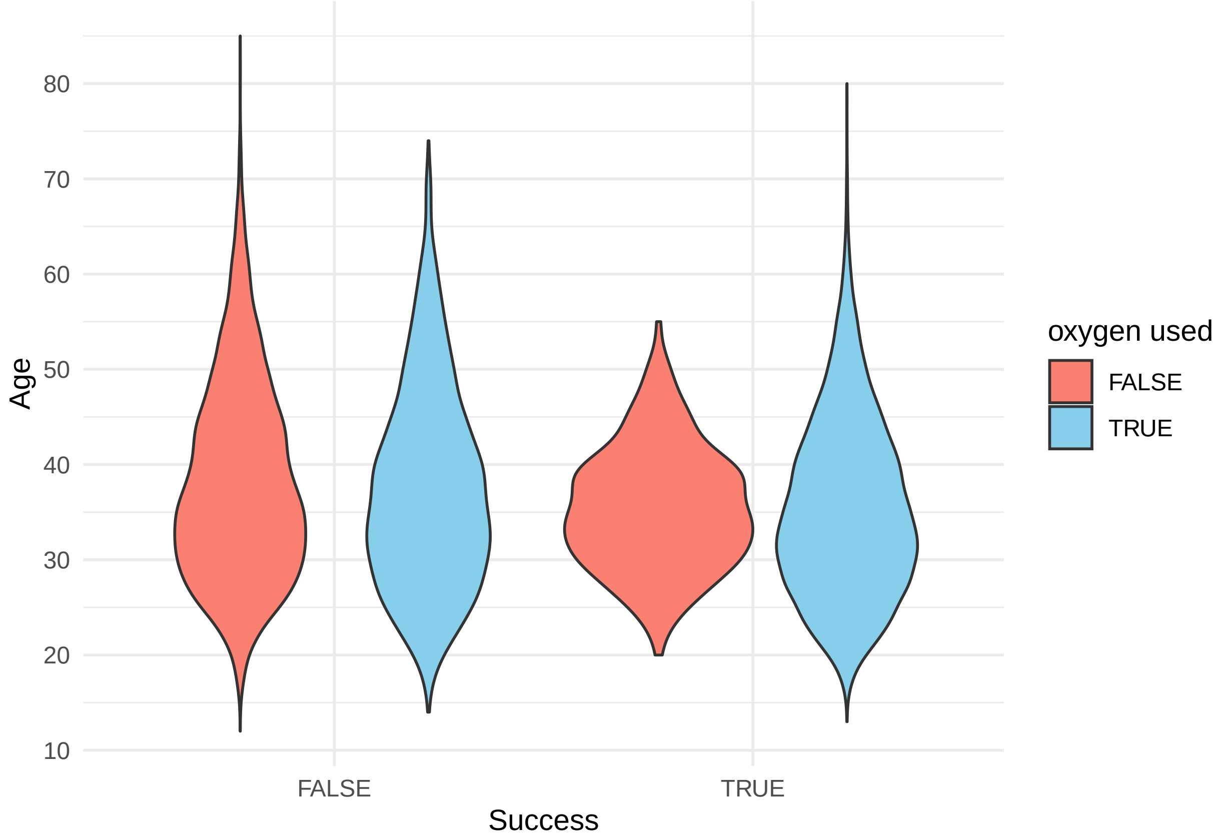
To answer the question of Part 1, we will work with four variables, the age of the individual who participated in the expedition (column age), whether the individual was successful in climbing a peak (column success), whether the individual used oxygen (column oxygen_used), and the year of expedition (column year).

Approach: *Our approach is to show the distributions of age versus successful/unsuccessful summit using violin plots (geom_violin()). We also separate out individuals who used oxygen and those who did not, as specified in the question. Violins make it easy to compare multiple distributions side-by-side.*

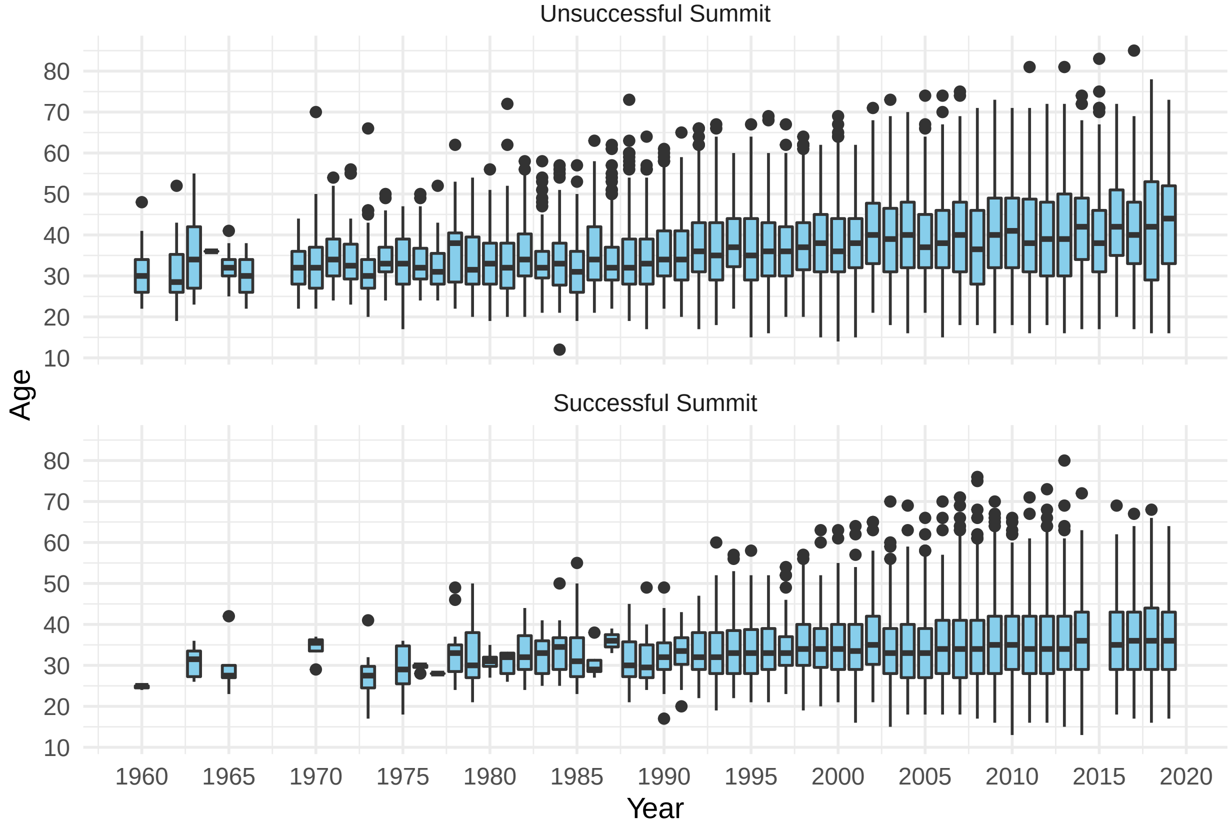
One limitation of the violin plots is that they don't show us how the age distribution changed over the years. Therefore, we will visualize the age distribution of different years with faceted boxplots (geom_boxplot()). Jointly, these two plots will allow us to answer the question.

Analysis:

```r
ggplot(members_everest, aes(x = success, y = age, fill = oxygen_used)) +
  geom_violin() + # create violin plot
  scale_y_continuous(breaks = seq(0, 100, by = 10)) + # set values for y axis
  scale_fill_manual(values = c(`TRUE` = "#87CEEB", `FALSE` = "#FA8072"),
                    name = 'oxygen used') + # assign color to violins
                                            # change the name of legend box
  # change names of x and y axis
  labs(x = 'Success', y = 'Age') +
  theme_minimal() # change theme
```



```r
ggplot(members_everest, aes(x = year, y = age, group = year)) +
  geom_boxplot(fill = 'skyblue') + # change the fill color
  # change the layout of the faceted plot
  facet_wrap(~ success, ncol = 1,
             labeller = as_labeller(c(`TRUE` = "Successful Summit",
                                      `FALSE` = "Unsuccessful Summit"))) +
  # change the values for x and y axis
  scale_x_continuous(breaks = seq(1960, 2020, by = 5)) +
  scale_y_continuous(breaks = seq(10, 90, by = 10)) +
  # change the name of x and y axis
  labs(x = 'Year', y = 'Age') +
  theme_minimal() # change theme
```



Discussion: *For expedition members who did not use oxygen, the age distributions appear to be different for members who were successful in climbing Mt. Everest and those who were not. We can see this by comparing the red violins in the first plot, where we see that the age distribution has a smaller variance for members who were successful. Successful or unsuccessful summit does not seem to affect the age distribution for people who used oxygen. This can be seen from the fact that the blue violins look approximately the same. For members who were unsuccessful in climbing Mt. Everest, the use of oxygen also does not affect the age distribution much because the red and blue violins on the left have similar shapes. The use of oxygen only affects the age distribution for successful members. The red violin on the right has a smaller variance compared to the blue violin. We would have to run a multivariate statistical analysis to determine whether any of these observed patterns are statistically significant.*

When we look at the age distribution for each year in the second plot, we see that the age distribution tends to have a higher median and a bigger variance as the year becomes more recent. The age distribution fluctuates more for unsuccessful members compared to successful members, where the median is stable since 1900. Thus, the final answer to the first question is: For members who were successful in climbing Mt. Everest, there are age differences for members who used oxygen and those who did not. For members who were not successful in climbing Mt. Everest, there are not much age differences for members who used oxygen and those who did not. Over the years, the age distribution tends to have a higher median and a bigger variance.

Part 2

Question: *Are there age differences for male and female expedition members who climbed Mt. Everest in different seasons and how does season affect the success in climbing Mt. Everest?*

Introduction: *We are also working with the* `members_everest` *dataset.*

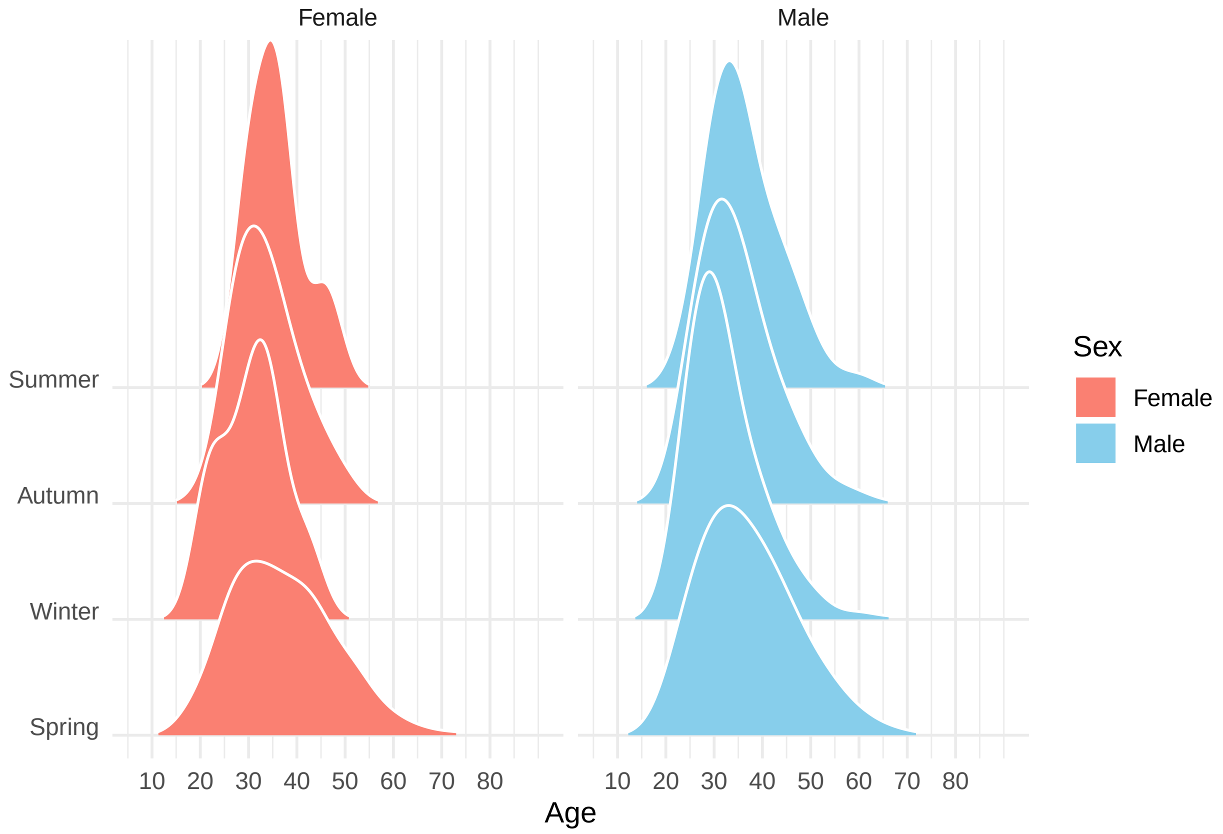
To answer the question of Part 2, we will work with four variables, the age of the individual who participated in the expedition (column age), the sex of the expedition member (column sex), whether the individual was successful in climbing to a peak (column success), and season of expedition (column season).

Approach: *Our approach is to show age distribution versus season using ridgeline plots (geom_density_ridges()). We also separate out male and female members, as specified in the question. Ridgeline plots make it easy to compare the shape of multiple distributions.*

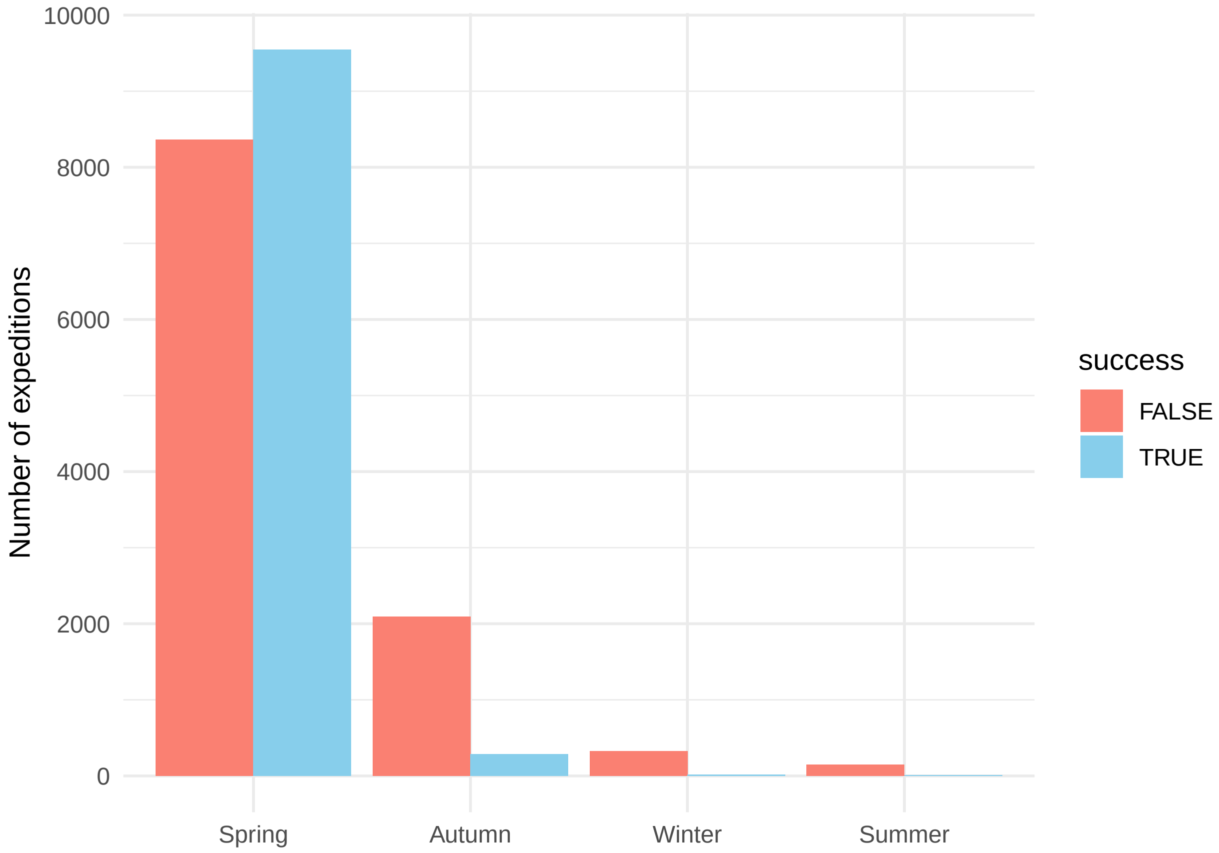
To explore the effect of season on the success in climbing Mt. Everest, we decide to show the number of expeditions for climbing Mt. Everest versus season using bar plots (geom_bar()). We also separate out female and male members, because sex might have different effects on them and should be considered separately. Dodged bar plots make it easy to compare number of expeditions side-by-side.

Analysis:

```r
ggplot(members_everest) +
  # change the order of season on y axis
  aes(x = age, y = fct_relevel(season, "Spring", "Winter", "Autumn", "Summer"),
      fill = sex
  ) +
  # create ridgeline plot
  geom_density_ridges(
    scale = 3, bandwidth = 3.4, rel_min_height = 0.01,
    color = "white"
  ) +
  # set fill colors and change name & labels of the legend box
  scale_fill_manual(values = c(`M` = "#87CEEB", `F` = "#FA8072"),
                    name = 'Sex', labels = c('Female', 'Male')) +
  # change the name and values for x axis and eliminate the gap
  scale_x_continuous(name = "Age", expand = c(0, 0),
                     breaks = seq(0, 80, by = 10)
  ) +
  # change the name of y axis and the gap
  scale_y_discrete(
    name = NULL,
    expand = expansion(add = c(0.2, 2.4))
  ) +
  # change theme
  theme_minimal() +
  # move the position of labels in y axis
  theme(axis.text.y = element_text(vjust = 0)) +
  facet_wrap(~sex, labeller = as_labeller(c(`F` = "Female",
                                      `M` = "Male"))
  )
```



```
ggplot(members_everest, aes(
  # change the order of seasons on x axis
  x = fct_relevel(season, "Spring", "Autumn", "Winter", "Summer"),
  fill = success)
  ) +
  # create bar plot
  geom_bar(position = 'dodge') +
  # change names of x and y axis
  labs(x = NULL, y = 'Number of expeditions') +
  # change the fill colors
  scale_fill_manual(values = c(`TRUE` = "#87CEEB", `FALSE` = "#FA8072")) +
  # change the values on y axis
  scale_y_continuous(breaks = seq(0, 10000, by = 2000)) +
  # change theme
  theme_minimal()
```



Discussion: *For both female and male expedition members, the age distributions are similar for members who climbed Mt. Everest in summer, autumn, and winter but different for members who climbed Mt. Everest in spring. We can see this by comparing the all the red ridgelines together and then comparing all the blue ridgelines together in the first plot, where we can see that the age distribution for spring has a higher variance and is more spread out compared to the age distributions for other seasons. The age distribution appear to be similar for female and male members for each season except for winter. In the first plot, the red and blue ridgelines are similar to each other for spring, summer, and autumn. For winter, male age distribution is more concentrated than female age distribution. We would have to run a multivariate statistical analysis to determine any of these observed patterns are statistically significant.*

When we look at the number of successful and unsuccessful expeditions in different seasons, we see that most of the expeditions take place in the spring. The majority of successful expeditions happens in spring, and the rest happens in autumn. There are no successful expedition in summer and winter. In spring, there are more successful expeditions than unsuccessful expeditions. We can see these by comparing the red and blue bars in the second plot. Thus, the final answer to the second question is: there are age differences for members who climbed Mt. Everest in spring compared to other seasons. For female and male members, there are age differences for those who climbed Mt. Everest in winter. Most of the successful expedition occur in spring and has the highest success rate.