Project 1

Samantha Uy Tesy

This is the dataset you will be working with:

```
members <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/20

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
```

```
as_tibble(members_everest)
```

```
## # A tibble: 20,790 x 21
##    expedition_id member_id peak_id peak_name  year season sex     age
##    <chr>         <chr>     <chr>   <chr>     <dbl> <chr>  <chr> <dbl>
##  1 EVER63101     EVER6310~ EVER    Everest    1963 Spring M        36
##  2 EVER63101     EVER6310~ EVER    Everest    1963 Spring M        31
##  3 EVER63101     EVER6310~ EVER    Everest    1963 Spring M        27
##  4 EVER63101     EVER6310~ EVER    Everest    1963 Spring M        26
##  5 EVER63101     EVER6310~ EVER    Everest    1963 Spring M        26
##  6 EVER63101     EVER6310~ EVER    Everest    1963 Spring M        29
##  7 EVER63101     EVER6310~ EVER    Everest    1963 Spring M        44
##  8 EVER63101     EVER6310~ EVER    Everest    1963 Spring M        37
##  9 EVER63101     EVER6310~ EVER    Everest    1963 Spring M        32
## 10 EVER63101     EVER6310~ EVER    Everest    1963 Spring M        26
## # ... with 20,780 more rows, and 13 more variables: citizenship <chr>,
## #   expedition_role <chr>, hired <lgl>, highpoint_metres <dbl>, success <lgl>,
## #   solo <lgl>, oxygen_used <lgl>, died <lgl>, death_cause <chr>,
## #   death_height_metres <dbl>, injured <lgl>, injury_type <chr>,
## #   injury_height_metres <dbl>
```

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?

To answer this question, we will plot the distribution of age by whether or not the climber used oxygen, and we will also plot the age distribution of climbers that were successful and not successful, respectively.

Introduction: We are working with the `members_everest` dataset, which contains 20,790 records over Mount Everest expeditions from 1960 through Spring 2019. In this dataset, each of the rows corresponds to a single expedition ID with 21 columns providing information about the expedition and member. Information about the expedition includes: expedition ID, peak name, peak ID, year, and season. Information about the member includes: member ID, sex, age, citizenship, expedition role, whether or not the member climbed solo, whether or not the member used oxygen, whether or not the member was hired, whether or not the member was injured, whether or not the member died in the expedition, and whether or not the expedition was successful. Subsequently, if the member was injured or died in the expedition, the information includes the cause of injury/death as well as the height at which the injury or death took place.

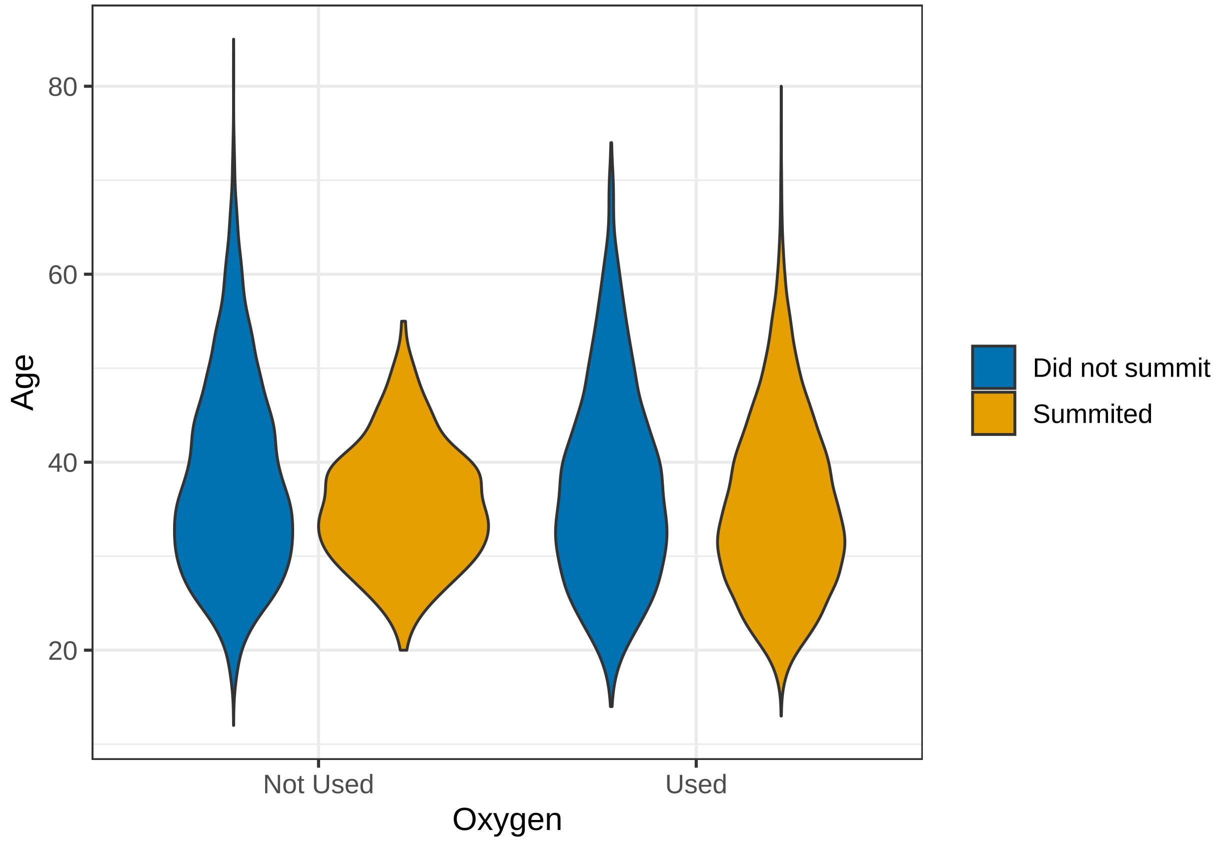
To answer the question in Part 1, we will work with four variables: the member's age (column `age`), whether or not the expedition was successful (column `success`), whether the member used oxygen (column `oxygen_used`), and the year of the expedition (column `year`). The age is provided as a numeric value, in years. The success status is encoded with the boolean variable TRUE/FALSE, where TRUE means the member was successful and FALSE means the member was not successful. The oxygen variable is encoded with the boolean variable TRUE/FALSE, where TRUE means the member used oxygen and FALSE means the member did not use oxygen. The year variable is provided as a numeric value, in years.

Approach: My approach is to show the distributions of age versus success status using violin plots (`geom_violin()`). I also separated the plots out by whether or not oxygen was used as it may affect the likelihood of success and must be considered separately. Violins make it easy to compare the distributions side by side.

A limitation of this approach is that the plots do not show how age distributions may have changed over time. Therefore, I will visualize the distribution of members' ages by year and separate them into two plots: "Successful" and "Not Successful", using (`geom_boxplot()`). Jointly, these two plots will allow us to ask the question in Part 1.

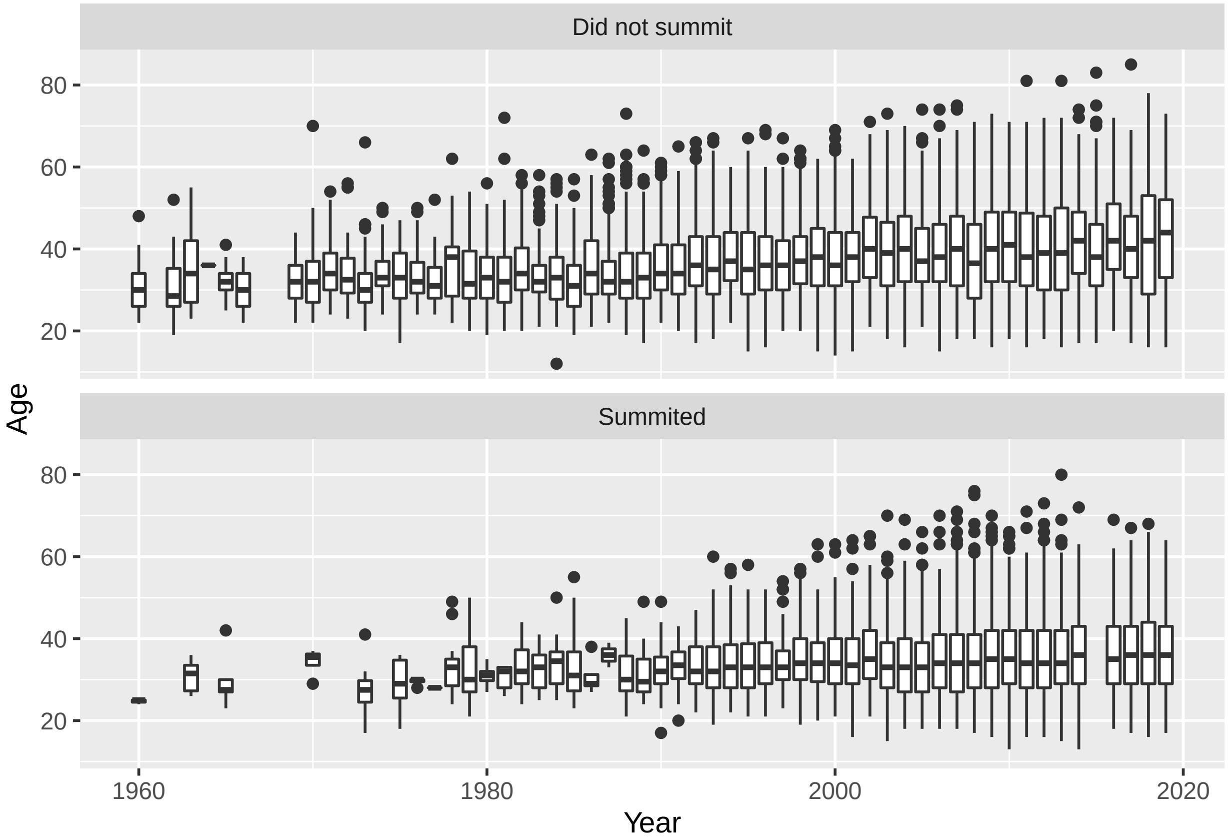
Analysis: First we plot the age distribution as violins.

```r
ggplot(members_everest, aes(x = oxygen_used ,y = age, fill = success)) +
  geom_violin() +
  # label the y-axis
  scale_y_continuous(
    name = "Age"
  ) +
  # label the x-axis and variables
  scale_x_discrete(
    name = "Oxygen",
    labels = c("Not Used", "Used")
  ) +
  scale_fill_manual(
    # label the legend
    name = NULL,
    labels = c(`FALSE` = "Did not summit", `TRUE` = "Summited"),
    # manually set colors of each plot
    values = c(`FALSE` = "#0072B2", `TRUE` = "#E69F00")
  ) +
  theme_bw(12)
```



Then we plot the distribution of ages as boxplots over time. We facet by whether or not the expedition was a success so that it is clear how the age distributions have changed in each subset of data.

```r
ggplot(members_everest, aes(group = year, year, age))+
  geom_boxplot() +
    facet_wrap(
       # separate data by summited and not summited
      vars(success),
       # organize facets into single column
      ncol = 1,
      labeller = as_labeller(
        c(`TRUE` = "Summited", `FALSE` = "Did not summit") # label the facets
      )
    ) +
  xlab("Year") + # label the x-axis
  scale_y_continuous( # label the y-axis
    name = "Age"
  )
```



Discussion: For those members who where successful, there was a larger distribution in the ages of the members in those who used oxygen. We can see this by comparing the yellow violins, where we see the age distribution for those who used oxygen and successfully summited has a larger range. For the members who were not successful, the age distributions are similar for those who used oxygen compared to those who did not. We can see this by comparing the dark blue violins, both of which have very similar shapes and little to no shift relative to one another. We would have to run a multivariate statistical analysis to determine whether any of these observed patterns are statistically significant.

When we add the additional dimension of time, we see that the age distributions have increased over time for both the “Successful” and “Not Successful” groups. The 1st and 3rd quartile range appears to have increased overtime for both groups. Thus, there appears to be no drastic difference between the age distributions of members to used oxygen compared to members that did not use oxygen. Furthermore, the age distribution has increased over time. We would have to run a multivariate statistical analysis to determine whether any of these observed patterns are statistically significant.

Part 2

Question: Is there a relationship between the success of an expedition and the season in which the expedition took place? Does the highest point reached in the expedition distribution change from season to season?

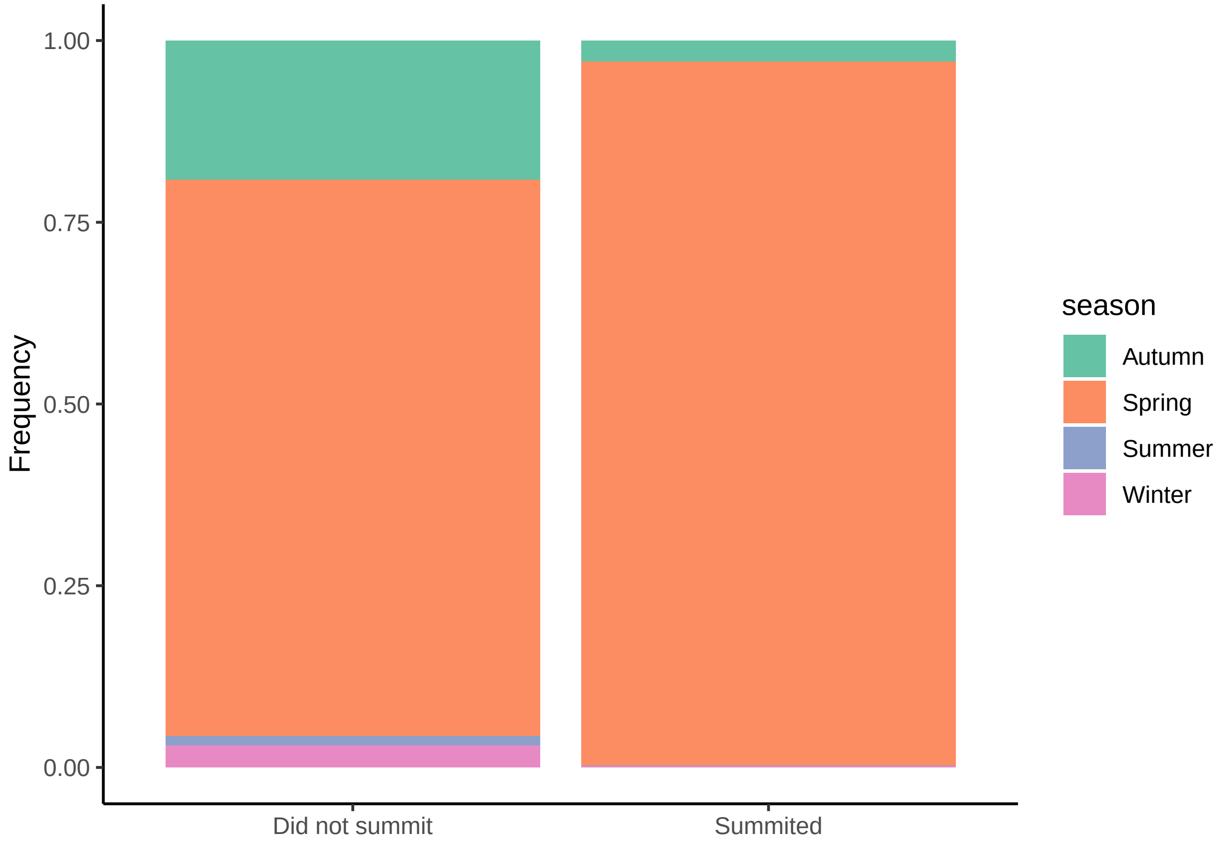
Introduction: To answer the question in Part 2, we will work with three variables: the season in which the expedition took place (column `season`), whether or not the expedition was successful (column `success`), and the high point (column `highpoint_metres`). The season status is provided as a qualitative variable, encoded as {Summer, Spring, Autumn, Winter}. The success status is encoded with the boolean variable TRUE/FALSE, where TRUE means the member was successful and FALSE means the member was not successful. The high point variable is provided as a numeric variable, in metres, and describes the highest point reached during the expedition.

Approach: Our approach is to show the relative frequency of members who summited versus members who did not summit Mount Everest by season using the `geom_bar()`bar plot.

One limitation of using bar graphs to answer this question is that, for those who did not summit, it cannot be determined whether the season is driving the result. Therefore, we will visualize the distribution of high points reached during each season using a ridgeline plot with `geom_density_ridges()`. Jointly, these two plots will allow us to answer this question.

Analysis: First we plot success by season in a bar graph.

```
ggplot(members_everest, aes(success, fill = season)) +
  # create bar plot with relative frequencies
  geom_bar(position = "fill") +
  scale_x_discrete(
    # label x-axis
    name = NULL,
    # label x-axis ticks
    labels = c(`FALSE` = "Did not summit", `TRUE` = "Summited"),
  ) +
  scale_y_continuous(
    # label y-axis
    name = "Frequency"
  ) +
  theme_classic() +
  # change color scheme to more muted color scheme
  scale_fill_brewer(palette = "Set2")
```

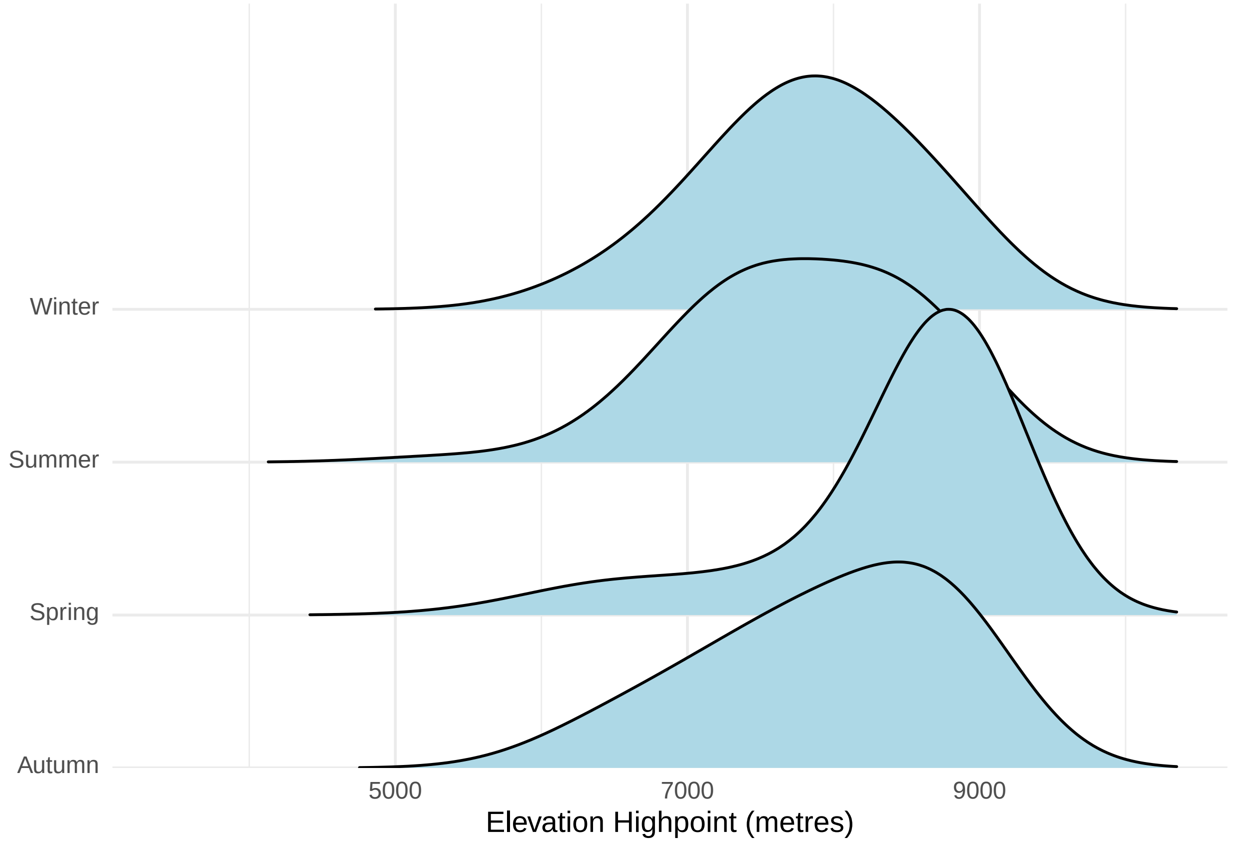


Then we plot the distribution of high points as densities on the ridgelines plot so we can clearly see how the distributions look in each subset of the data.

```r
# import geom_density_ridges from the ggridges library
library(ggridges)

# this plot answers the question, what is the distribution of height in each season
ggplot(members_everest, aes(x = highpoint_metres, y =season)) +
  geom_density_ridges(
    # color density functions
    fill = "lightblue",
    # fit the density functions onto the graph
    scale = 2, bandwidth = 500,
    rel_min_height = 0.001
  ) +
  scale_y_discrete(
    name = NULL,
    # fixing the width of y-axis to fit
    expand = expansion(add = c(0, 2))
  ) +
  scale_x_continuous(
    # labeling the x-axis
    name = "Elevation Highpoint (metres)"
  ) +
  theme_minimal() +
  theme(
    # centering the y-axis ticks
    axis.text.y = element_text(vjust = 0.3)
  )
```

```
## Warning: Removed 5001 rows containing non-finite values (stat_density_ridges).
```



```
# 5001 rows removed for containing N/A value for hightpoint_metres
```

Discussion: For those who did not summit, it appears that the majority of expeditions took place in the Autumn and Spring seasons, with very few expeditions taking place in the Summer or Winter season. For those who summited, it appears that even more of the expeditions took place in the Spring season, with few expeditions in Autumn, and fewer if not zero expeditions in the Summer or Winter season.

When we look at the density functions in the ridgelines plot, we see that the median high point reached in the Spring is relatively higher compared Winter and Summer. Similarly, the median high point in Autumn is relatively higher compared to Winter and Summer, although, not as extreme as in the Spring. Therefore, it makes more sense to compare the high point reached by season to determine whether the season in which the expedition took place has any effect on the climb. It is important to note that 5001 expeditions are not present in the ridgelines plot because there is no recorded high point value.

Thus,the season appears to have an effect on the success of a climb by a large amount. We would have to run a multivariate statistical analysis to determine whether any of these observed patterns are statistically significant.