Project 1

Derian Lee (dtl634)

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: The *members* dataset is taken from the Himalayan Database, and for this project, we are working with a subset of that data (*members_everest*) which contains information about expeditions to Mount Everest since 1960, along with the known ages of the adventurers. The question to answer is (1) whether there are differences in age distributions when looking at the usage of oxygen and success of the climber, and (2) whether or not the distribution of ages has changed over time for those who did and did not successfully climb. To answer this question, several variables from the dataset will be needed. The following variables will be taken from the *members_everest* dataset: *age, oxygen_used, success,* and *year.*

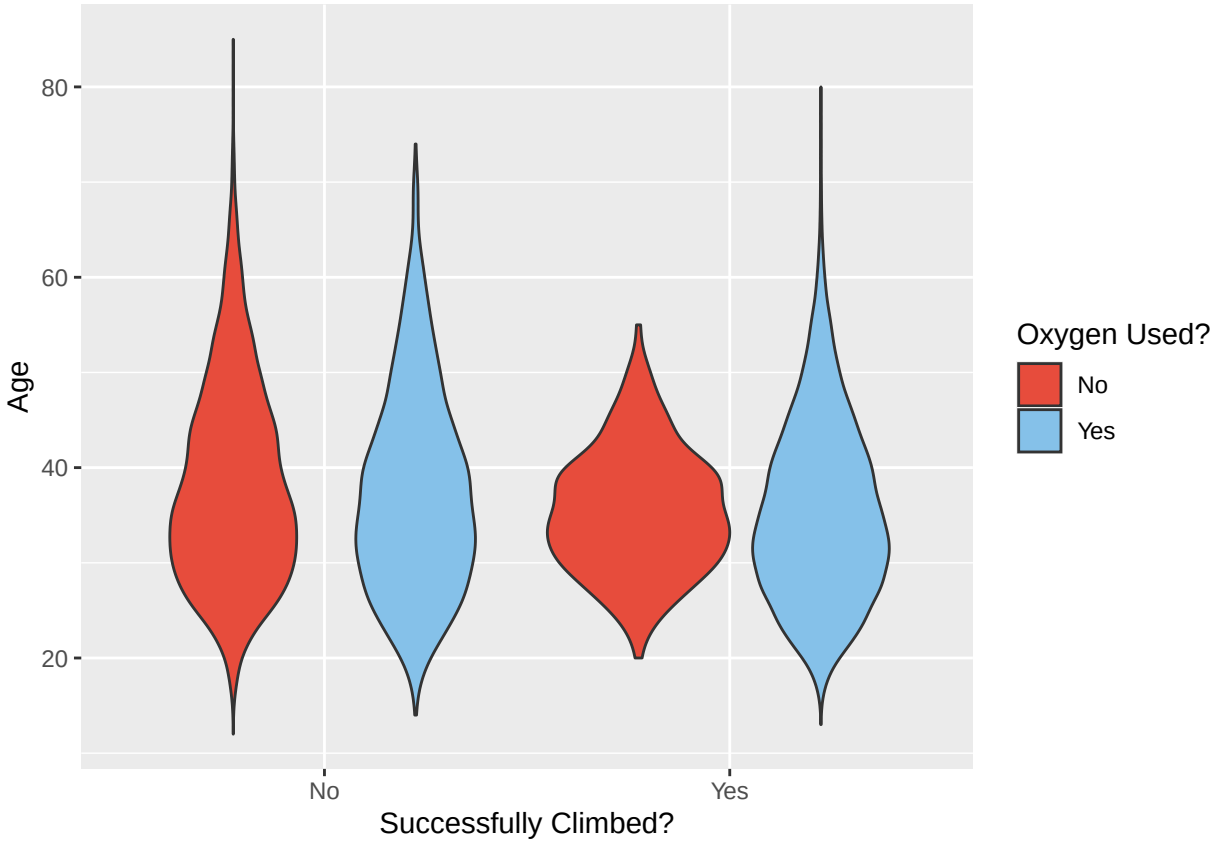
Approach:

For the first part of the question, a violin plot will be used. A violin plot was chosen because it can show density distributions of a numeric variable (age) for more than one category (oxygen usage and success). Furthermore, it is a good choice because the categories in question are T/F categories, meaning our graph will only have 4 violins (distributions will be easier to compare with fewer violins).

For the second part of the question, a faceted boxplot will be used. This was chosen because faceted boxplots allow us to visualize the spread of a numeric variable (age) over a categorical variable (year), then facet across the two groups (success). This is a solid choice because boxplots are a quick way to communicate visual information, and can thus be used over a large range of values (having many boxes) without being difficult to interpret. For example, it is simple to answer questions like how the median age has changed throughout the years, simply by looking at how the middle line of each box moves up or down when reading the boxes from left to right.

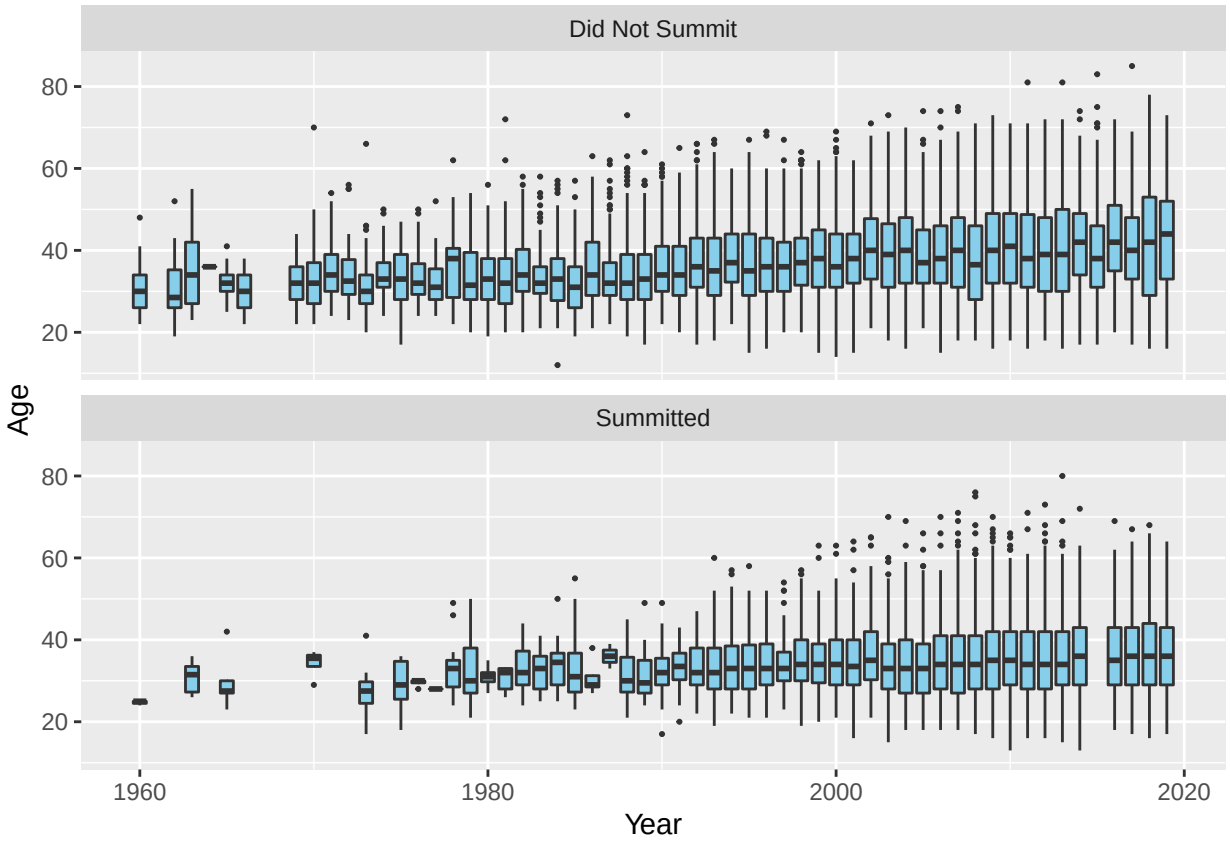
Analysis:

```
ggplot(members_everest, aes(x = success, y = age, fill = factor(oxygen_used))) + #setting axes
  geom_violin() +
  scale_x_discrete( #renaming the x axis title, along with the label names
    name = "Successfully Climbed?",
    labels = c("No", "Yes")
  ) +
  scale_y_continuous( #renaming the y axis title
    name = "Age"
  ) +
  scale_fill_manual( #renaming the legend title and names, coloring the key
    name = "Oxygen Used?", labels = c('FALSE' = "No", 'TRUE' = "Yes"),
    values = c(`FALSE` = "#E74C3C", `TRUE` = "#85C1E9"))
```




```
ggplot(members_everest, aes(x = year, y = age, group = year)) + #setting axes, formatting
  geom_boxplot(fill = "skyblue", outlier.size = 0.4) +
  labs(x = "Year", y ="Age") +
  facet_wrap(vars(success), #separate by success
```

```
ncol =1, labeller = as_labeller(
c(`TRUE` = "Summitted", `FALSE` = "Did Not Summit")) #renaming 'success' arguments
)
```



Discussion:

First graph: For those who failed in their climb, the age distribution of those who used oxygen has a smaller age spread, but overall has similar aspects to those who did not use oxygen, visually speaking. For those who succeeded in their climb, the age distribution for those who used oxygen has a noticeably wider age spread compared to those who didn't, especially towards the higher end of the age spectrum. In other words, as age increases, it becomes increasingly more difficult to climb Everest successfully without oxygen. This makes sense, as people tend to lose stamina the older they get.

Second graph: *FAILURE GROUP:* Over the years, the median and IQR of age distributions of those who failed their climb has been steadily increasing, but the rate of change is low. Throughout the years, it's become more difficult for people to successfully climb Everest as they age, but the span of ages that have attempted to climb Everest annually has widened. This may be due to technological advancements, which may encourage more people to climb Everest, even if they are not in the best shape (knowing they have back-up oxygen, emergency kits might encourage a wider audience to try). *SUCCESS GROUP:* Over the years, the median age of people who succeeded in climbing Everest has remained relatively constant, while the IQR has more or less stabilized starting around year 2000. The flip side of technological advancements is that it makes it easier for a wider range of ages to succeed as well (and remain stable over time). Perhaps Everest became more commercialized during this time and expedition groups became more popular, improving group protocol, instruction, and overall solidarity.

Part 2

Question: When using oxygen, how does probability of death differ for the sexes, and what other variables in the dataset may be informative to reduce death rate?

Introduction: For Part 2, the *members_everest* dataset will be used again (information about the dataset described above). The question to answer is (1) when using oxygen, how does probability of death differ for the sexes, and (2) what other variables within the dataset may be responsible for death. To answer this question, the following variables will be chosen and examined from the *members_everest dataset: oxygen_used, sex, season, highpoint_metres,* and *died.*

Approach:

For the first part of the question, we will use a grouped bar plot. By using the *(position = "fill")* argument, we can scale the dead / alive observations onto a scale of 1, allowing us to visualize the probabilities of death across different groups (*oxygen_used*). Using a grouped bar plot makes it easy to compare probabilities – one only needs to compare the horizontal widths of the boxes in question, side by side. The use of contrasting colors (red/blue) further accentuates this difference, making the comparison very easy. The bar plot is then faceted by sex, offering a comprehensive graph of how death varies across oxygen usage for both males and females.

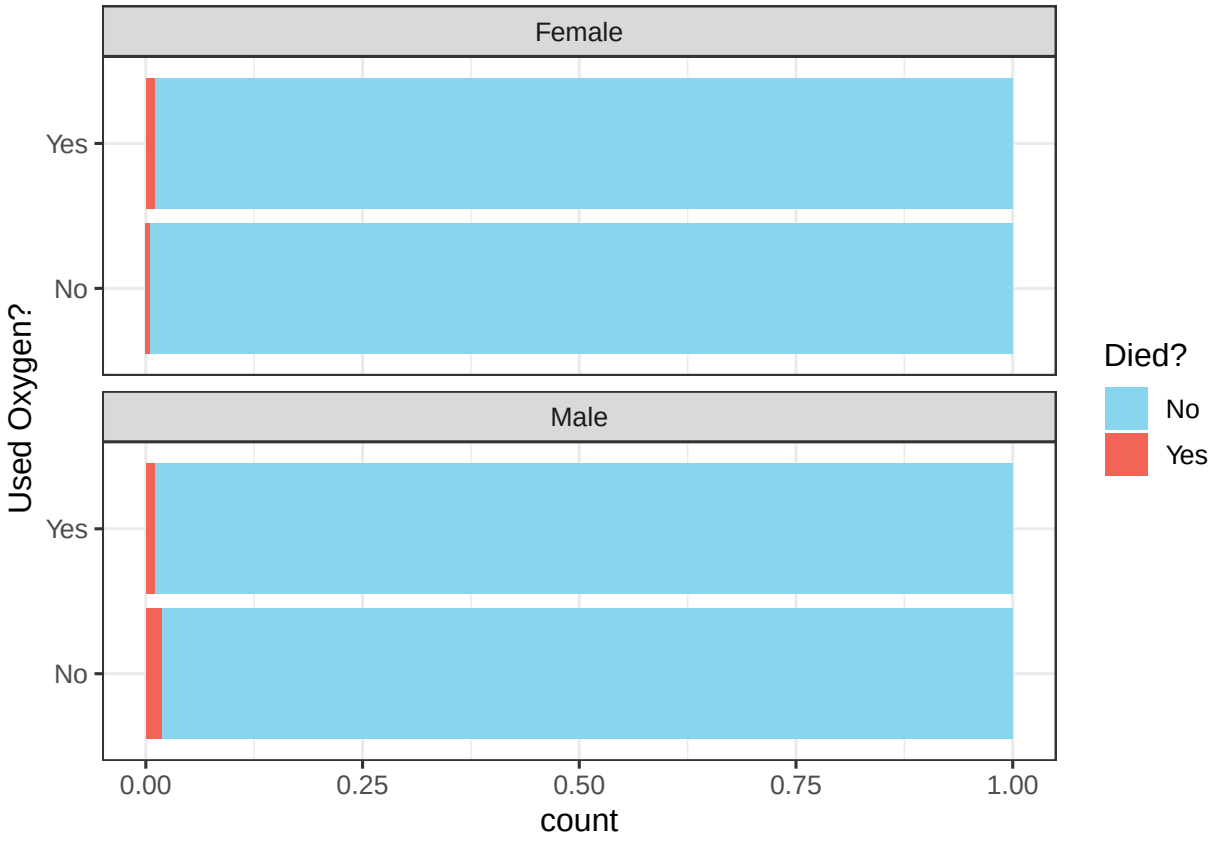
For the second part of the question, we will look at the effects of (a) *season* and (b) *highpoint_metres* on death distributions. Because there are null values in the *members_everest* dataset for *highpoint_metres*, the data will be filtered to remove empty highpoint values. Ridgeline plots are similar to violin plots in that they can be used to show density distributions of a numeric variable (in this case, *highpoint_metres*) for more than one categorical variable (*season* and *died*). The reason why a ridgeline plot was chosen is because the categorical variable *season* has 4 groups. If we were to make a violin plot, plotting living status across season would subsequently create 8 violins, side-by-side. Thus, this choice was made due to aesthetic reasons: overlaying plots reduces visual stimuli and consequently, improves interpretability.

Analysis:

```
table(members_everest$sex) #returns count for males and females
```

```
##
##     F     M
##  1702 19088
```

```
ggplot(members_everest, aes(y = oxygen_used, fill = died)) + #setting axes
  geom_bar(position = "fill" #scaling bars to be the same lengths
  ) +
  facet_wrap( #separating by sex
    vars(sex),
    ncol = 1,
    labeller = as_labeller(c(`F` = "Female", `M` = "Male"))
  ) +
  scale_y_discrete( #setting labels
    name = "Used Oxygen?",
    limits = c("FALSE", "TRUE"),
    labels = c("No", "Yes")
  ) +
  scale_fill_manual( #setting labels, colors
    name = "Died?",
    labels = c('FALSE' = "No", 'TRUE' = "Yes"),
    values = c(`FALSE` = "#8AD5EE", `TRUE` = "#F16456")
  ) +
  theme_bw(12)
```



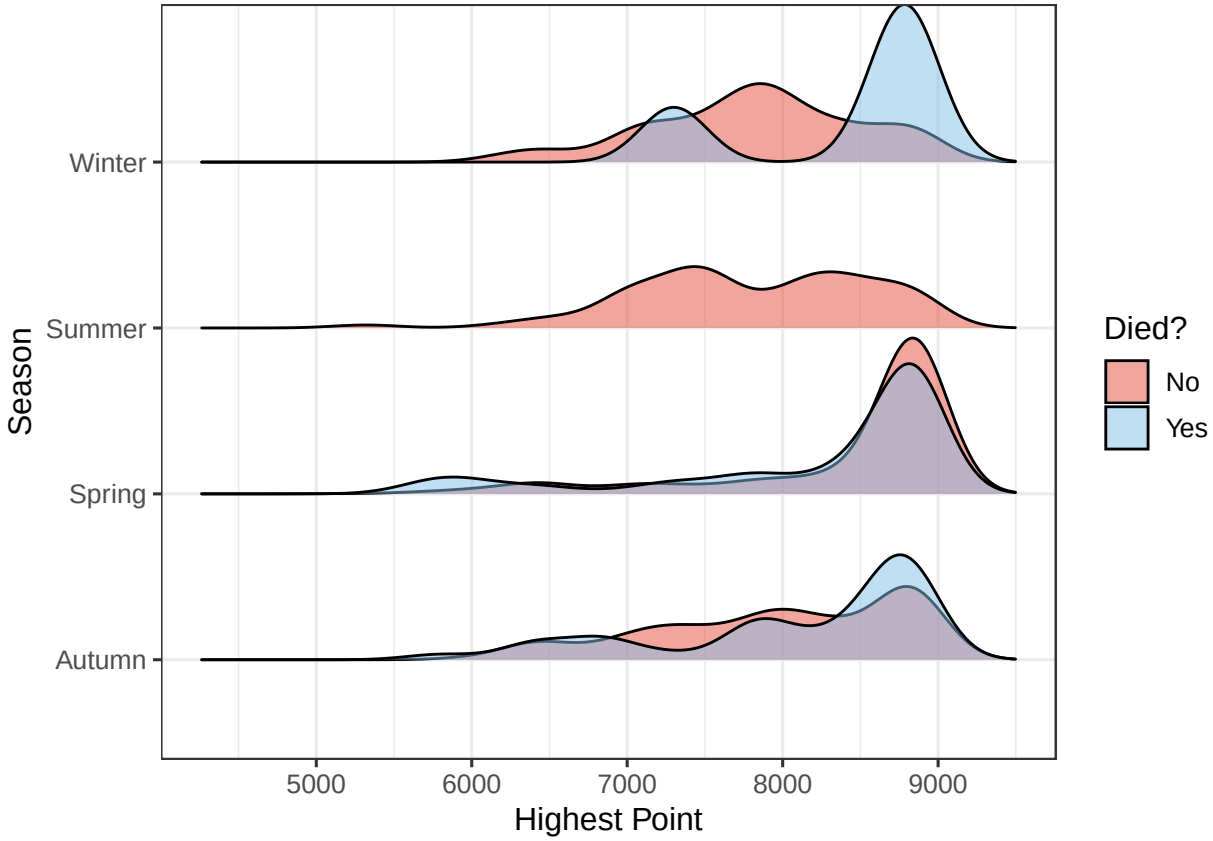

```
library(ggforce)
library(ggridges)

members_everest2 <- members_everest %>% #filtering data
  filter(!is.na(highpoint_metres))

table(members_everest2$season, members_everest2$died) #freq table for season / death
```

```
##
##          FALSE  TRUE
##   Autumn  1105    41
##   Spring 14294   152
##   Summer    74     0
##   Winter   119     4
```

```
ggplot(members_everest2, aes(x = highpoint_metres, y = season, fill = factor(died))) + #setting axes
  geom_density_ridges(alpha = 0.5, scale = 0.95) + #setting transparency and distance between plots
  scale_x_continuous(
    name = "Highest Point"
  ) +
  scale_y_discrete(
    name = "Season"
  ) +
  scale_fill_manual(
    name = "Died?", labels = c('FALSE' = "No", 'TRUE' = "Yes"), #renaming, coloring legend
    values = c(`FALSE` = "#E74C3C", `TRUE` = "#85C1E9")
  ) +
```

```
  theme_bw(12)
```

```
## Picking joint bandwidth of 216
```



Discussion:

First graph: Visually speaking, males who do not use oxygen are about twice as likely to die than males who do use oxygen. Interestingly, the inverse is true for females: Females who use oxygen are about twice as likely to die than females who do not use oxygen. When comparing between groups, males and females are about equally likely to die when using oxygen, but males are more prone to dying than females when not using oxygen. When comparing between groups, a limitation is that the sample size for males and females widely differ, so these visual conclusions may not be reliable. Though increased survival was expected when using oxygen, this was not the case for females. A speculative explanation could be that males overestimate their athletic ability (in this case, ability to climb without oxygen) while females have a more realistic ability to gauge their ability. In other words, male misperception may explain why they are more likely to die in the absence of oxygen. A final limitation to this graph and subsequently, this discussion, is that the proportion of those who died across all groups is so small that it is difficult to make precise measurements (ie. hard to determine "group A is more likely to die by a factor of x compared to group B").

Second graph: During the winter and autumn seasons, people who reached extreme heights (8500m - 9000m) were more likely to die than survive. Thus, as a word of caution, future climbers should be wary of climbing to extreme heights during either of these seasons. During spring, those who died compared to those who survived had very similar highpoint distributions. Interestingly, people who climbed extreme heights during this season were more likely to survive than die. Perhaps one possible reason is due to group mentality and encouragement: many more people climb Everest on Spring than any other season and thus, larger groups may foster motivation. During the summer, nobody died. Though the sample size for summer is too small to infer, it would make sense intuitively that the warm temperatures of summer would make it one of

the safer seasons to climb. A similar limitation is seen for the winter season, where the small sample size for winter observations is likely responsible for the strange bimodal distribution. In conclusion, *highpoint_metres* may be a helpful indicator in predicting death, but the extent of its effect on death seems to vary across seasons, and a larger sample size (as well as statistical testing) would be needed to confirm any interaction.