Project 1

Xiangrui Pan

XP742

Data Visualization

3/8/2021

This is the dataset you will be working with:

```r
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
```

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

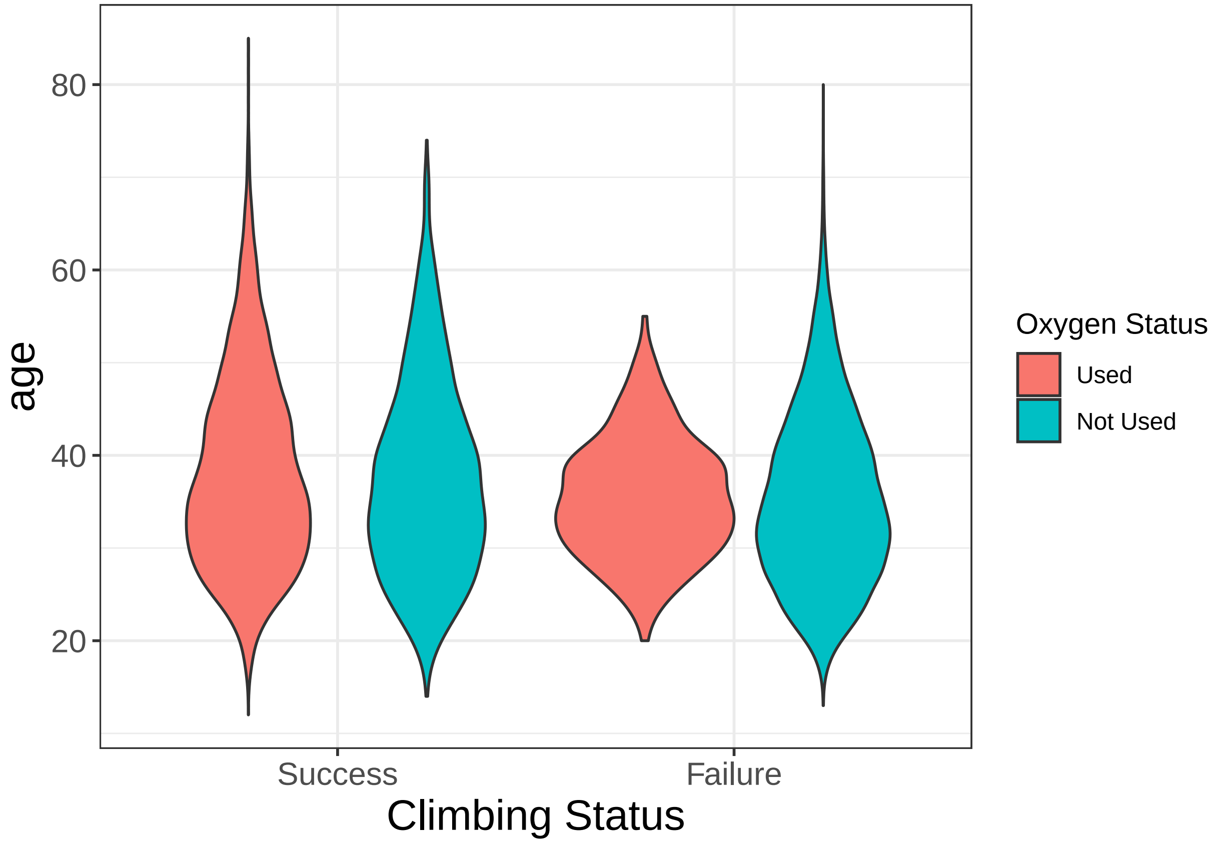
Question: Are there age differences for expedition members who were successful or not in climbing Mt. Everest with or without oxygen, and how has the age distribution changed over the years?

Introduction: We will be working with Himalayan dataset called `members_everest` which has the records of expedition to Nepal Himalaya from year 1960 to Spring 2019. In this dataset, we can see members who climbed Himalayan usually come as a group, that is, several `member_id` correspond to the same `expedition_id`. There are more than 10 columns which provide us information about the trips. We are interested in whether or not age could be a factor that expedition members who climbed Mt. Everest with or without oxygen and the change of age distribution over the years. To answer the question above, we actually do not need to use all the column that provided. Instead, we will be using several explantory variables that include `age`(the age when members climbed Mt. Everest), `success`(to check if the person was successfully climbed), `oxygen_used`(if the person used oxygen during the trip), `year` (in what year they climbed). Notice that `success` and `oxygen_used` are encoded as two levels qualitative variables, take the value either TRUE or FALSE, TRUE means the member successfully completed climbing/ has used oxygen and FALSE means did not finish the expedition due to some reason/did not use oxygen. Also, `age` and `year` are encoded as quantitative variables, in years.

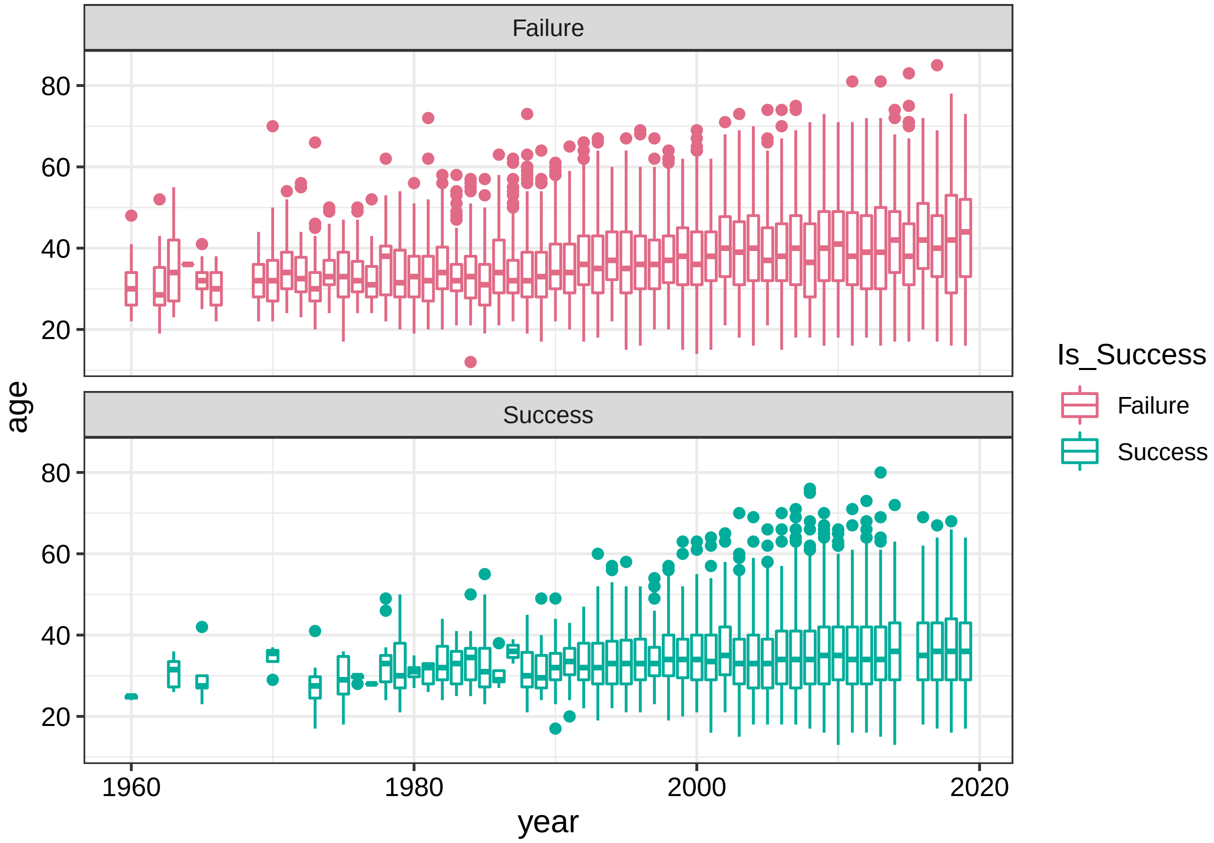
Approach: Our approach to the first part of the question is to use violin plot. We will use violin plot to show the age versus the climbing status (success or failure), and then I will fill oxygen status to each plot. The reason I choose violin plot because it can show us the distribution of the data, so we can visualize the shape of about what age that has higher chance to climb Mt. Everest successfully. And for the second part of question, I decide to use faceted boxplots to see how the age changes will affect the climbing status over the years. Since we need to compare two groups (success and failure), faceted function can help us to separate these two groups. Using boxplots because boxplots can provide median value. In our context, the median value is median age of climbers. Also, boxplots can show the maximum and minimum value and some outliers to let us have the some insights.

Analysis:

```
#using violin plot as suggested
ggplot(members_everest, aes(x = factor(success), y = age, fill = factor(oxygen_used))) +
         geom_violin()+
  #b/c x is categorical variable
  scale_x_discrete(name = "Climbing Status", labels = c("Success","Failure")) +
  #using "fill" here b/c we chose to "fill" oxygen_used earilier.
  scale_fill_discrete(name = "Oxygen Status", labels = c("Used", "Not Used")) +
  theme_bw() +
   #using theme to make the plots look nicer
  theme(axis.title = element_text(size = 16), axis.text = element_text(size = 12))
```



```
library(colorspace)
#using boxplot us suggested
ggplot(members_everest, aes(x = year, y = age, group = year, color = success)) +
  geom_boxplot() +
  #using facet function to separate success and failure
  facet_wrap(vars(success),
             ncol = 1,
             labeller = as_labeller(c('TRUE' = "Success", 'FALSE' = "Failure"))) +
  theme_bw() +
  theme(axis.text = element_text(size = 10, color = "black"),
      axis.title = element_text(size = 12)) +
  scale_color_discrete_qualitative(name = "Is_Success", labels = c("Failure", "Success"))
```



Discussion: Looking at the first graph, we can easily answer the question that was asked earlier. Yes, there are age differences for expedition members who were successful or not in climbing Mt. Everest, but almost no difference with or without oxygen. As we can see that for not using the oxygen supply, the difference between the age is not very significant no matter they climbed it successfully or not by comparing the light blue violin shape. On the other hand, people who used the oxygen had significantly age differences. People who successful climbed and used the oxygen had more variance in age whereas the age for people who failed is mostly between 30 to 40-year-old. We can observe some interesting facts in this graph. For instance, age is not the main factor that made people failed to climb since there are a few climbers were already more than 60 when they successfully climbed the mountain. And, why there are more people between age 30 to 40 have failed to climb with or without using oxygen supply? One assumption to this question is since they were relatively younger, so they carried heavier equips which made them more vulnerable than the elders.

In the second graph, there is no significant change of ages in successfully climbing the mountain. We can see that the successes and failures seem to be the same after the year of 1990 and the age may differ a little but not too much. But there is a huge difference before 1980. There was only handful people who climbed the mountain successfully before 1980. I think that is because the climbing gears were not good enough compared with those made after 1980. Moreover, between year 1960 and 1990, age is one of the main factors that if the climbers can make to Mt. Everest. We can see from the graph that the climbers who made it had smaller variance in age, they were all around 20- to 40-year-old with only a few outliers. Whereas, for the climbers who did not made it, they had larger variance in age with more outliers. In the end, no matter the success or failure, all the climbers seem to have the approximately same median age around 30.

Part 2

Question: Are the injured condition differ based on the climbers ages and genders, and how the mortality status change over the years?

Introduction: For the second part of project, We will stick with the same dataset as I used in Part 1. Here, We will investigate into the injured condition for the climbers based on their ages and genders. I would make the assumption that the climbers who had injured tend to be male. And for the second part of the question, I am curious about whether or not the number of death decreases with the increases in years. My assumption to this part is there were less rate of death in 2020 than in 1960 To answer the question above. We would introduce some new column from the dataset. We will use the columns called `age`, `year` as before, and we will have some new variables called `sex`, `injured`, `died`. They are all categorical variables with two levels. `sex` gives us the information about the gender of a climber who is either male or female. `injured` provides us the information about the injured condition, True means they got injured during the trip whereas False means the climbers were not got hurt. Lastly, `died` simply tells us whether or not a climber died during the trip.

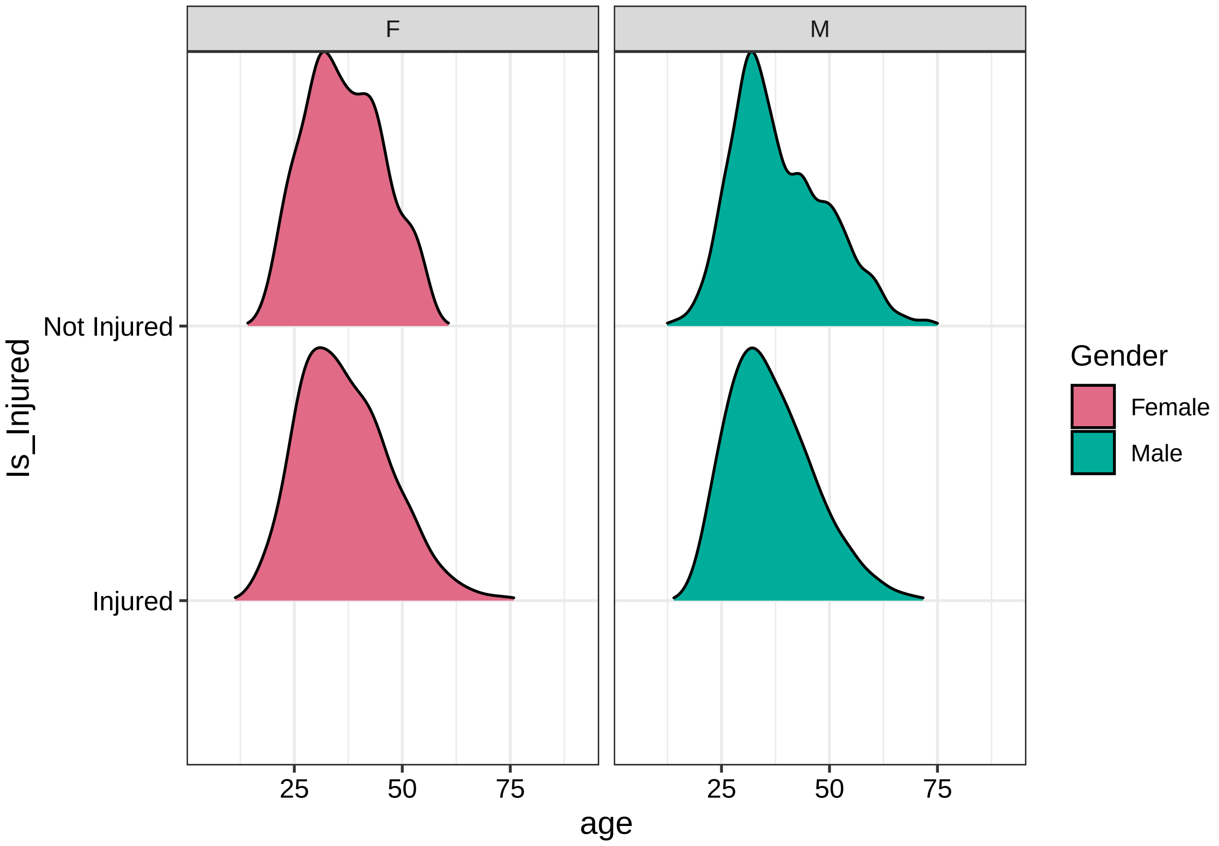
Approach: For the first part of this question, we are interested in whether the climbers got hurt based on their ages and genders. My approach to this part of the question is to use rigidline because rigidline is able to shows us the non-overlapped graph after setting the appropriate scale and using facets, so we can make the comparison much easier. My approach to the second part of the question is to use sina plots. Just like violin plot, sina plots can also show us the distribution of the variable but consists of many points. In this part, we are interested in the death of the climbers as the year progresses. We want to see the change in terms of shape, so I choose to use sina plots.

Analysis:

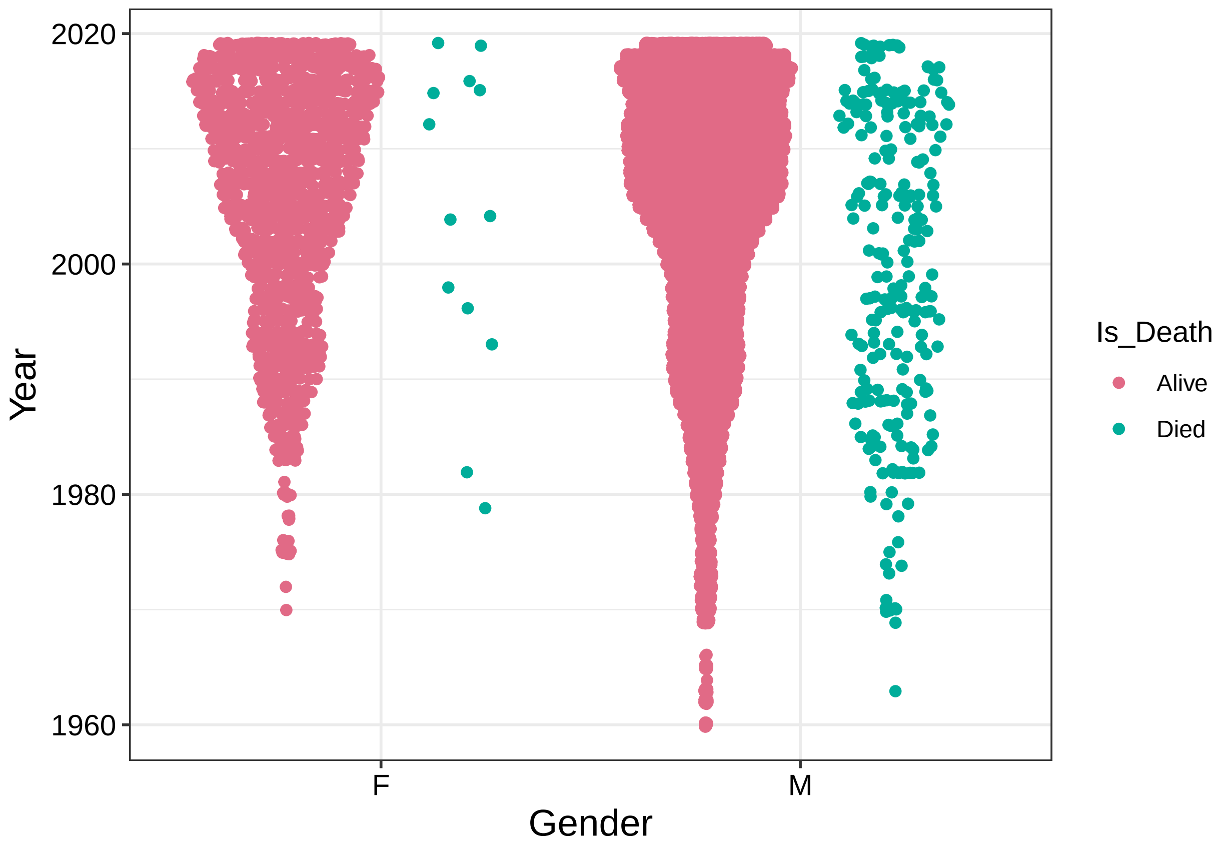
```r
library(colorspace)
library(ggridges) #we need to install library ggridges in order to use ridgeline
ggplot(members_everest, aes(x = age, y = factor(injured),  fill = factor(sex))) +
  geom_density_ridges(scale = 1, rel_min_height = 0.01 ) +
  #set scale = 1 to make the plot non-overlap, use rel_min_height = 1 to cut off extended lines
  facet_wrap(vars(sex)) +
  scale_y_discrete(name = "Is_Injured", labels = c("Injured", "Not Injured")) +
  #since I chose to use fill in aes function
  scale_fill_discrete_qualitative(name = "Gender", labels = c("Female", "Male")) +

  theme_bw()+
  theme(axis.text = element_text(size = 10, color = "black"),
        axis.title = element_text(size = 12, color = "black"))
```

```
## Picking joint bandwidth of 2.87
```

```
## Picking joint bandwidth of 2.08
```



```
# colorspace library is for scale function
library(colorspace)
# to use sina, we need to install the ggforce library
library(ggforce)
ggplot(members_everest, aes(x= factor(sex), y = year, color = factor(died))) +
  geom_sina() +
    theme_bw() +
  theme( axis.text  = element_text(size = 11, color = "black"),
         axis.title = element_text(size = 14, color = "black"),
         )+
  #using scale function to rename the variable
  scale_color_discrete_qualitative(name = "Is_Death", labels = c("Alive", "Died")) +
  scale_x_discrete(name = "Gender") +  #since x is discrete
  scale_y_continuous(name = "Year") #since y is discrete
```



Discussion: We can see that there is almost no difference in injured based on the climbers' genders and age from the first graph since the two distributions are almost identical, so we conclude that for both male and female climbers, the proportion of being injured is about the same at all ages. However, it is easy to observe a big difference when we switch our focus to the climbers who were not injured. Female climbers had more proportion of not getting injured in the age between 25 to 50 than male climbers. Since the ridgeline function only shows us the distribution of proportion, so we cannot conclude there were fewer female climbers got hurt than male climbers in terms of number of people. However, we can say female climbers were less likely to get hurt. In fact, the number of male climbers is 10 times more than the number of female climbers.

In the second graph, we want to see how the change of death according to the years. Clearly, no matter what year it was, the number of people alive is far greater than the number of deaths. The distribution for alive of both male and female is like funnel-shape through the years. That is because there were less people attempting to climb Mt. Everest in the early time and the climbing got more popular as the time approaching today. If we then compare the number of deaths to number of alive for both male and female, we can say the death rate is very low and we can run the statistical test to show the number of deaths are decreasing.