Project 1

Cristian Sigala crs4565

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: Today we'll be using the members dataset in order to explore some data about Mt.Everest. We'll be trying to see if there is any difference in age between expedition members who were successful or not on their climb depending on if they also used oxygen masks. Our second question is how has age distribution changed over the years in these groups. For both of these problems we will be looking at these variables "age" which is the age of the climber in years, "success" which indicates if they were able to climb to the peak or subpeak zone, and "oxygen_used" to see if the climber used any type of oxygen mask to assist their climbing. The variables "success" and "oxygen_used" are encoded in TRUTH/FALSE.

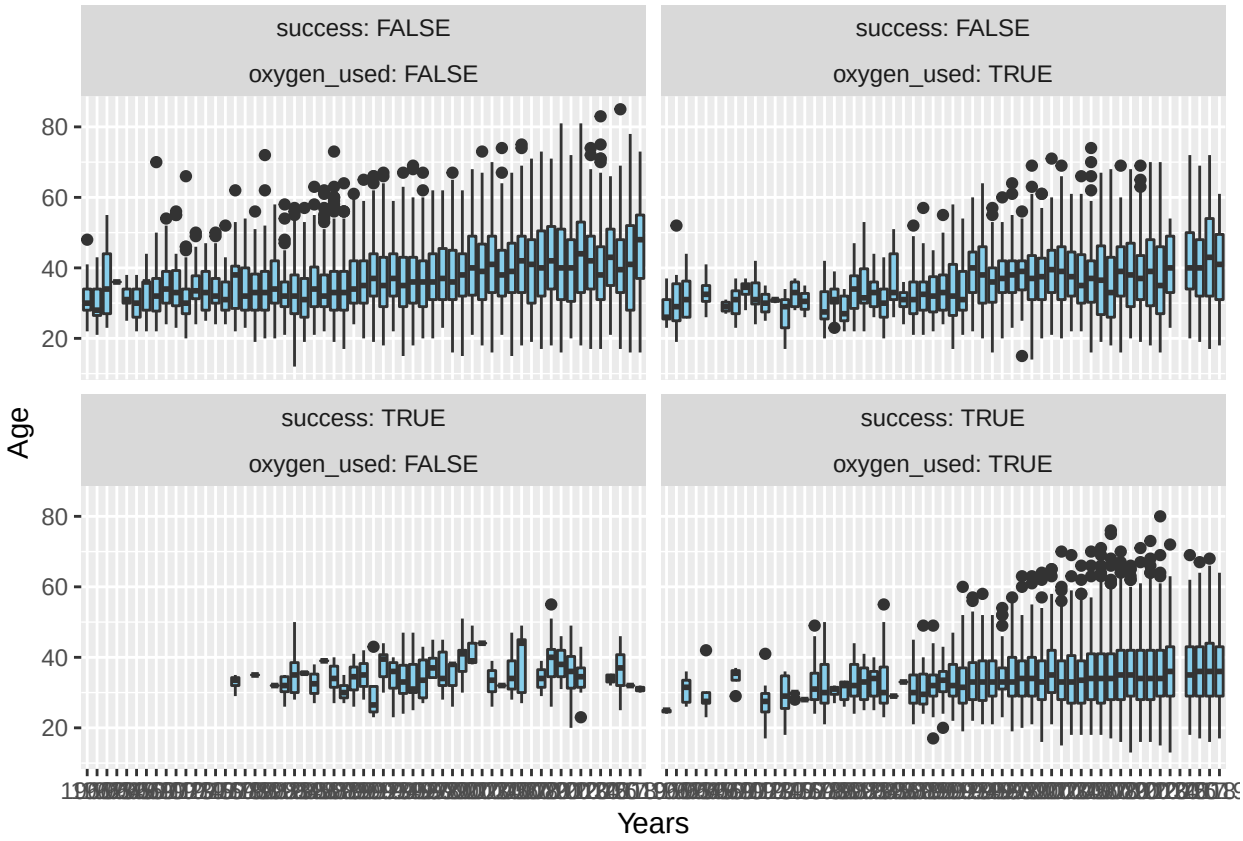
Approach: Before we can start looking at the data, it must be wrangled in order to only have data on Mt.Everest. We renamed this dataset to members_everest. For this first question we will be trying to see the distribution between three discrete categorical variables (oxygen_used , success , and age) so we will use violin plots. Violin plots are the most effective to see if there are any age differences between these groups because it shows the full distribution of data while still separating the variables side to side.

For the second question we will be using a boxplots to compare these groups over time. Boxplots are used because it gives us the ability to see the summary statistics of age distribution per year and compare those over time. We will use the facet function in order to isolate the variables and see exactly how the age distribution differs per condition.

Analysis:

```
ggplot(members_everest, aes(x = success , y = age , fill=oxygen_used)) +
  geom_violin() +
  scale_fill_manual(values=c("#FAD7A0", "#ABEBC6")) + #change the color of the violin plots
  xlab("Success") + ylab("Age") + # change the axis titles
  ggtitle("Age Distribution of climbers on Mt.Everest") +theme_classic() # Change title and theme
```


```
ggplot(members_everest, aes(x = factor(year),y= age)) + geom_boxplot(fill = "skyblue") +
  facet_wrap(success ~ oxygen_used , labeller=label_both) +   #Creating facet
  xlab("Years") + ylab("Age") #changing axis titles
```



Discussion: After constructing our violin plot, distribution between all these groups is large enough to see a visual difference. The first trend I noticed is that there is a big difference in range when climbers use oxygen masks vs when they are not. For example, when oxygen was not used the age range of successful climbers was significantly lower than those who use an oxygen mask. There are plenty of relationships to see from this plot, but the most important is there is definitely a significant difference in age range when looking at those who fail or succeeded depending if they use an oxygen mask or not.

For the second question, when isolating the interaction between the variables and plotting them over time, we can see different trends on all the graphs. When looking at the graphs, most of the groups see a steady increase in age distribution over the years. However, those who successfully climbed Mt.Everest without an oxygen mask have no trend but just sporadic distribution over time. This can be explained because we can only do so much to improve the ability of our body. However, with advancements in technology, oxygen masks are able to better perform allowing a wider range of people attempting to climb Mt.Everest.

Part 2

Question: Are there any differences for the amount of attempts and successes for expeditions during the different seasons, and how do these counts change when isolating for those who survived?

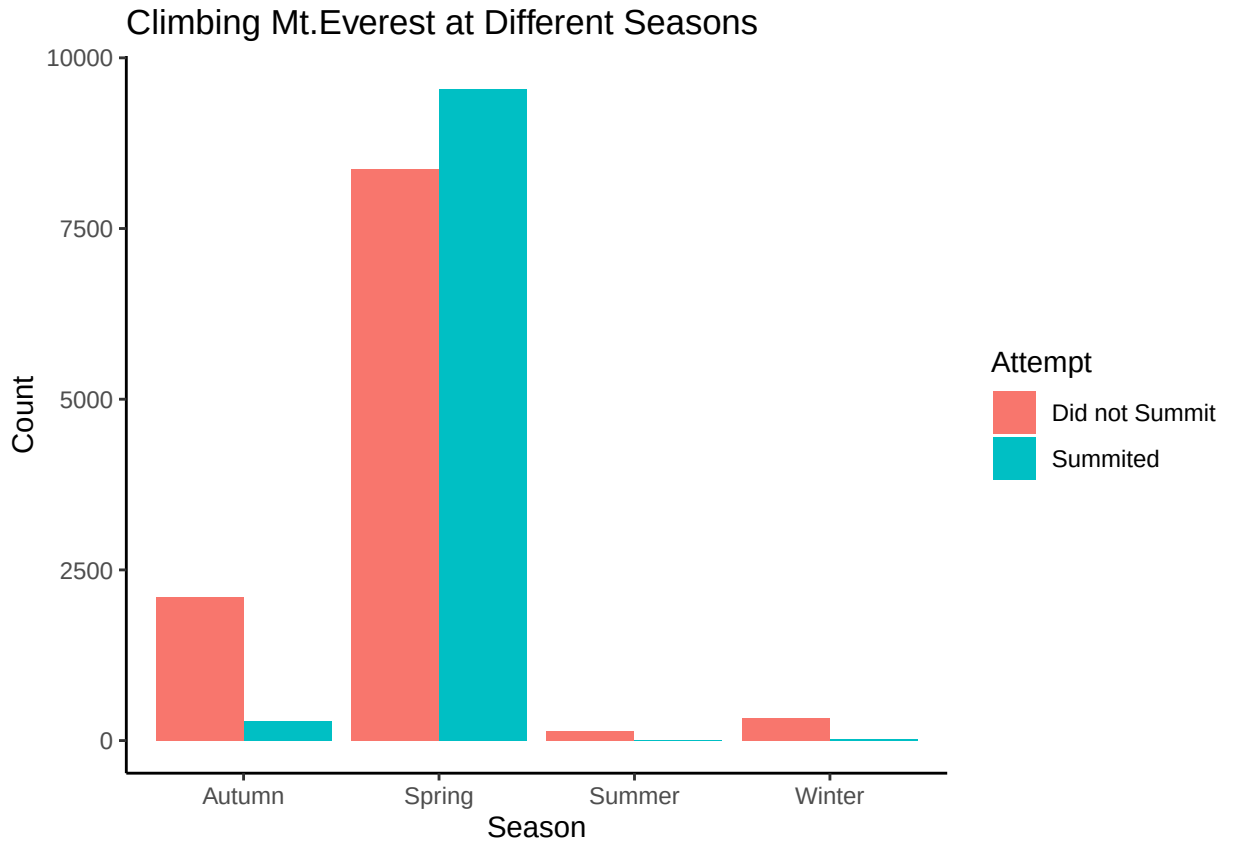
Introduction: While still using the members_everest dataset, we will now use "success", "seasons", and "died" to find any relationship between them. For the first part of the question we will be working with 2 different variables, success and season. Variable "Success" determines if the participants were able to reach peak or sub-peak. Variable "seasons" is just the season in which participants had the expedition. For our second question, we will reuse these same variables but use "died" as an additional variable. Variable "died" tells us if the participant died during their expedition. Variables "success" and "died" are encoded in TRUTH/FALSE.

Approach: For the first part of the question, all our variables are categorical (success, and season). Since we are interested to see the raw counts we will use a bar graph. A bar graph gives us a good way to compare the success and failures of different seasons. Depending on the difference of counts we can see if there is any relationship between these two variables.

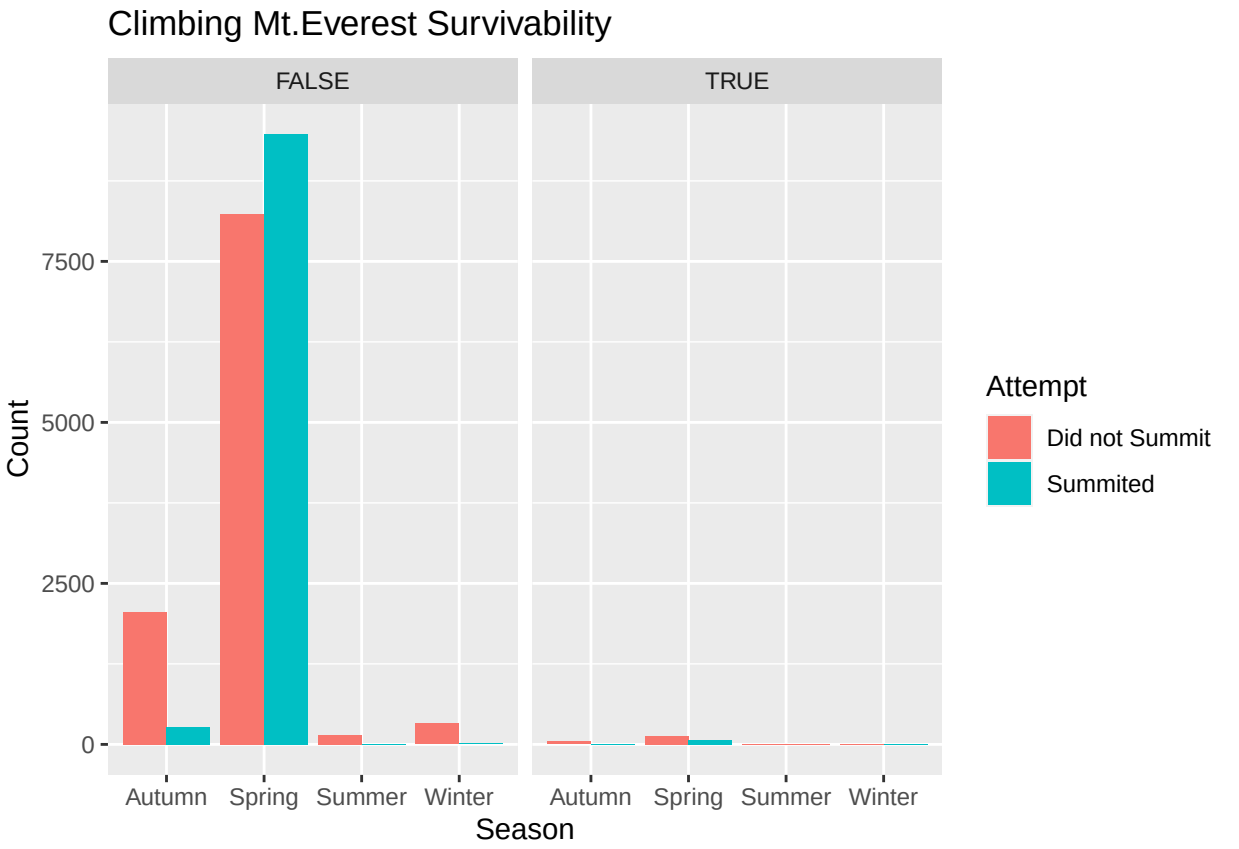
For our second part of the question, we will be conducting the same bar graph but faceting over the "died" variable. This will allow us to continue to see these raw counts while also seeing the distribution for those who survived vs those that died.

Analysis:

```
ggplot(members_everest, aes(x=season, fill = success)) + geom_bar(position = "dodge") +
  xlab("Season") + ylab("Count") + #Change axis titles
  ggtitle("Climbing Mt.Everest at Different Seasons") + #Change title
  scale_fill_discrete(name = "Attempt", labels = c("Did not Summit", "Summited")) +
  theme_classic()
```




```
ggplot(members_everest, aes(x=season, fill = success)) + geom_bar(position = "dodge") +
  facet_wrap(~died) + #creating facet around "died"
  xlab("Season") + ylab("Count") + #Changed axis titles
  ggtitle("Climbing Mt.Everest Survivability") + #Changed title
  scale_fill_discrete(name = "Attempt", labels = c("Did not Summit", "Summited"))
```




```
#Changed legend labels
```

Discussion: After making a boxplot we can see the difference in attempts between the seasons and how many people succeed in their climb. Initially I thought there would be an extremely low amount of attempts in the winter and autumn because the warmer weather would provide easier conditions to climb. To my surprise, there is a massive difference between the seasons with most attempts being held during Autumn and Spring. After some research I found that during the summer it is monsoon season which creates this mist that covers the mountain top. I also found that most attempts are maid during the month of May because those are the best conditions to climb. That is why we see such a large difference in attempts during the spring vs all the other seasons.

For the second question, when isolating those who survived, we can see some interesting trends. The first thing that I noticed is that the amount of deaths on this mountain is surprisingly low considering all the hazards. Another is that most people die during the spring but that is expected due to the sheer amount of people attempting the climb. Lastly,I noticed that there was still deaths even in those who succeeded in climbing the mountain. The decent is obviously just as dangerous but often overlooked.