Project 1

Sookja Kang, sk26949

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"))
  )
```

```
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>
```

```
summary(members_everest)
```

```
##  expedition_id       member_id           peak_id           peak_name
##  Length:20790       Length:20790       Length:20790       Length:20790
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##       year         season              sex                 age
##  Min.   :1960   Length:20790       Length:20790       Min.   :12.00
##  1st Qu.:1997   Class :character   Class :character   1st Qu.:29.00
##  Median :2007   Mode  :character   Mode  :character   Median :35.00
##  Mean   :2004                                         Mean   :36.68
##  3rd Qu.:2014                                         3rd Qu.:43.00
##  Max.   :2019                                         Max.   :85.00
##
##  citizenship        expedition_role      hired         highpoint_metres
##  Length:20790       Length:20790       Mode :logical   Min.   :4910
##  Class :character   Class :character   FALSE:14493     1st Qu.:8250
##  Mode  :character   Mode  :character   TRUE :6297      Median :8850
##                                                       Mean   :8394
##                                                       3rd Qu.:8850
##                                                       Max.   :8850
##                                                       NA's   :5001
##   success           solo          oxygen_used        died
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:10933     FALSE:20783     FALSE:8490      FALSE:20529
##  TRUE :9857      TRUE :7         TRUE :12300     TRUE :261
##
##
##
##
##  death_cause        death_height_metres  injured        injury_type
##  Length:20790       Min.   :4600        Mode :logical   Length:20790
##  Class :character   1st Qu.:6000        FALSE:20315     Class :character
##  Mode  :character   Median :7850        TRUE :475       Mode  :character
##                     Mean   :7368
##                     3rd Qu.:8500
##                     Max.   :8830
##                     NA's   :20532
##  injury_height_metres
##  Min.   :4000
##  1st Qu.:6956
##  Median :8000
##  Mean   :7675
##  3rd Qu.:8700
```

```
##  Max.   :8880
##  NA's   :20498
```

Introduction:

I am using 'members_everest' dataset that contains 20790 expedition members at Mt. Everest from 1960 to 2019. In this dataset, each row corresponds to one expedition member, and there are 21 columns providing detailed information about the expedition members and climbing. The information about expedition members includes unique identifier for expedition, unique identifier for the person, sex of the person, age of the person, citizenship of the person, role of the person on the expedition, whether the person was hired. The information about climbing contains unique identifier for peak, peak name, expedition year, season of expedition, elevation highpoint of the person, whether the person was successful in summitting a peak, whether the person attempted a solo ascent, whether the person used oxygen, whether the person died, primary cause of death, height at which the person died, whether the person was injured, primary cause of injury, and height at which the injury occurred.

To answer question 1 of the first part, I will work with 3 variables, the person's age (age), success in summitting a peak or not (success), and used oxygen or not (oxygen_used). The age is provided as a numeric value. The status of summitting a peak is encoded as TRUE or FALSE, where TRUE means success and FALSE means fail. Also, the status of using oxygen is encoded as TRUE or FALSE, where TRUE means use of oxygen and FALSE means no use of oxygen. For question 2 of the first part, I will use the 3 variables, the person's age (age), year of expedition (year), and success in summitting a peak or not (success). The age and the status of summitting a peak are the same values as the first part. The year is provided as a numeric value.

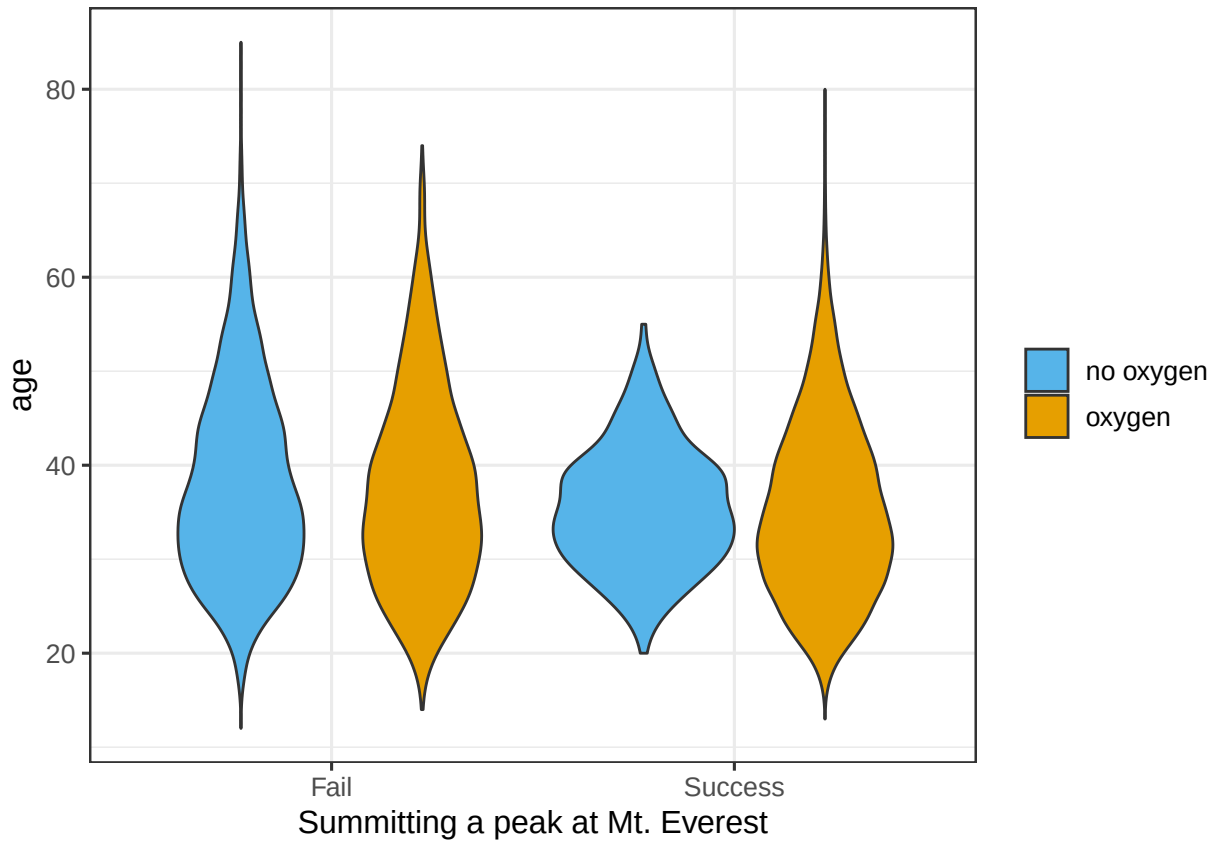
Approach:

During this analysis, I am going to use two plots to answer questions. The first approach is to show the distributions of ages versus the status of summiting a peak using violin plots (geom_violin()). I also separate out the use of oxygen and no use of oxygen to show how the use of oxygen affects the status of summiting a peak. With these violin plots, it is easy to compare distribution across different categories side by side. These violin plots have a limitation of not showing a total number of observations for each category. The second approach is to visualize the distributions of ages versus different years using box plots (geom_boxplot()). I will facet (facet_wrap()) by the status of summiting a peak. This will make it easy to compare age differences across the different years. Also, the facet function will allow me to make a comparison of the age difference between the two groups, success in summiting a peak or not. This boxplot also has a limitation that does not show the number of observations for each year.

Analysis:

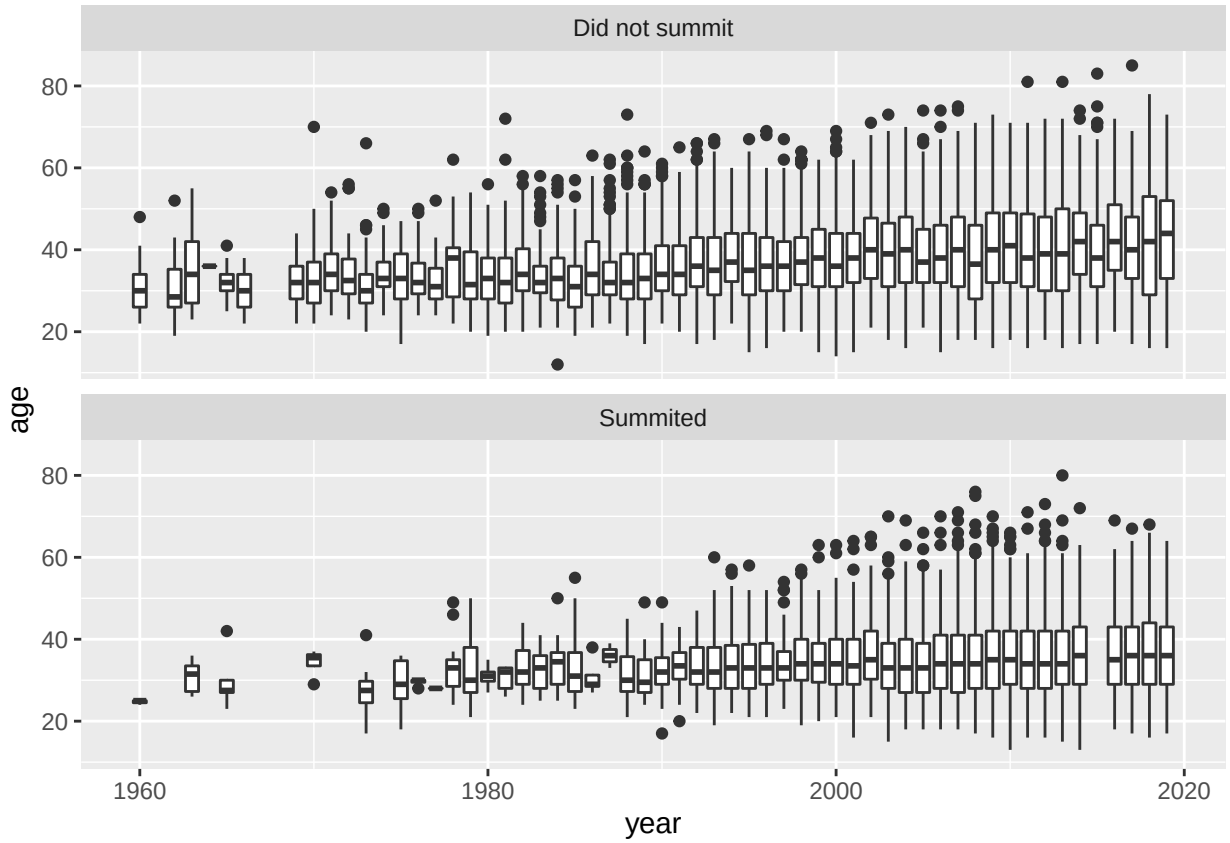
I plot the age distribution using violins.

```
# Your R code here
ggplot(members_everest, aes(factor(success), y = age, fill = oxygen_used)) +
  geom_violin() +
  scale_x_discrete(
    name = "Summitting a peak at Mt. Everest",
    labels = c ("Fail", "Success")
  ) +
  scale_fill_manual(
    name = NULL,
    labels = c("no oxygen", "oxygen"),
    values = c('FALSE' = "#56B4E9", 'TRUE' = "#E69F00")
  ) +
  theme_bw(12)
```



Then I plot the age changes over years using box plots. I facet by the summitting status of Mt. Everest whether the person was successful in summiting a peak or not. This provides a better understanding of age changes over years in each subset.

```
# Your R code here
ggplot(members_everest, aes(year, age, group = year)) +
  geom_boxplot() +
  facet_wrap(vars(success),
             ncol = 1,
             labeller = as_labeller(c('TRUE' = "Summited", 'FALSE' = "Did not summit")))
```



Discussion:

For no use of oxygen, the success in summiting a peak at Mt. Everest appears to have a narrow range of the age distribution (20-60) compared to the failure in summiting a peak. We can see this by comparing the blue violins. With the use of oxygen, age distributions across the status of summiting a peak do not show wide differences. When we compare the orange violins, the age range of the success in summiting a peak has slightly wider than the one in the fail in summiting a peak. In the group of successfully summiting a peak, the use of oxygen has a much wider age range than no use of oxygen (on the right side of the plot chart). In the group of failed summiting a peak, no use of oxygen has a wider age distribution than the use of oxygen. This shows that expedition members in the elderly population can have a large negative outcome in summiting a peak without the use of oxygen.

When we look at the age distribution across the different years from 1960 to 2019 by summited/not summited, we see wider age ranges in the not summited group compared to one in the summited group. The outliers in the not summited group have higher ages than ones in the summited groups. Also, median values in the summited group tend to lower than ones in the not summited group. In both plots, the age ranges in the later years show more variances than ones in the beginning. Thus, I can conclude that there are changes in the age distribution over the years.

Part 2

Question:

1) Are there high point differences between the summited group and not summited group across different gender groups?

2) What are the top 5 countries of expedition members' citizenship? Are there differences in the numbers of

the expedition members across different seasons?

Introduction:

I am using the same dataset, including 20790 expedition members and 21 variables, that I used for part 1. There are no modifications in the 21 variables. The detailed descriptions of the variables are included in the introduction of part 1. To answer the first question of part 2, I will work with 3 variables, the elevation highpoint of the person (highpoint_metres), summited a peak or not (success), and the person's sex (sex). The elevation highpoint of the person is a numeric variable. The status of summiting a peak is encoded as TRUE or FALSE, where TRUE means summited and FALSE means not summited. The sex is encoded as F or M, where F means female and M means male. To answer the second For the second question of part 2, I will use the 2 variables, citizenship of the person (citizenship) and season of the expedition (season).

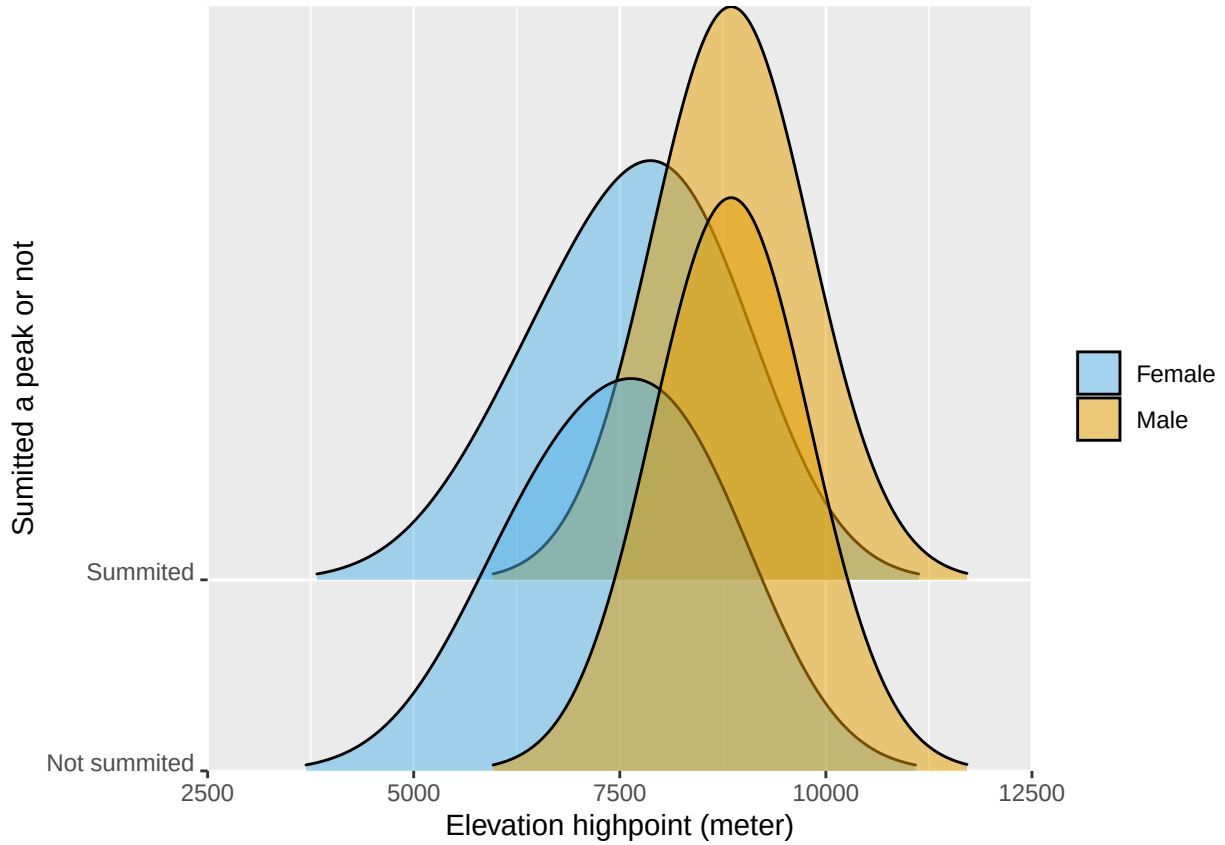
Approach:

My approach for question 1 is to visualize the distribution of the elevation highpoint in the two groups, 1) summited a peak and 2) not summited a peak, using the ridge plots (geom_density_ridges()). I also separate out female and male members to describe the gender difference on the elevation highpoint distribution. These ridge plots allow me to easily compare the elevation highpoint across both the gender types and the summited status. However, these plots do not provide the total numbers of observations that fall into each category.
The approach for question 2 is to show 5 countries with the highest numbers of expedition members categorized by citizenship using the bar plots (geom_bar()). Also, I will facet (facet_wrap()) by four seasons. This will make it easy to compare the changes in the numbers of expedition members across the different seasons.

Analysis:

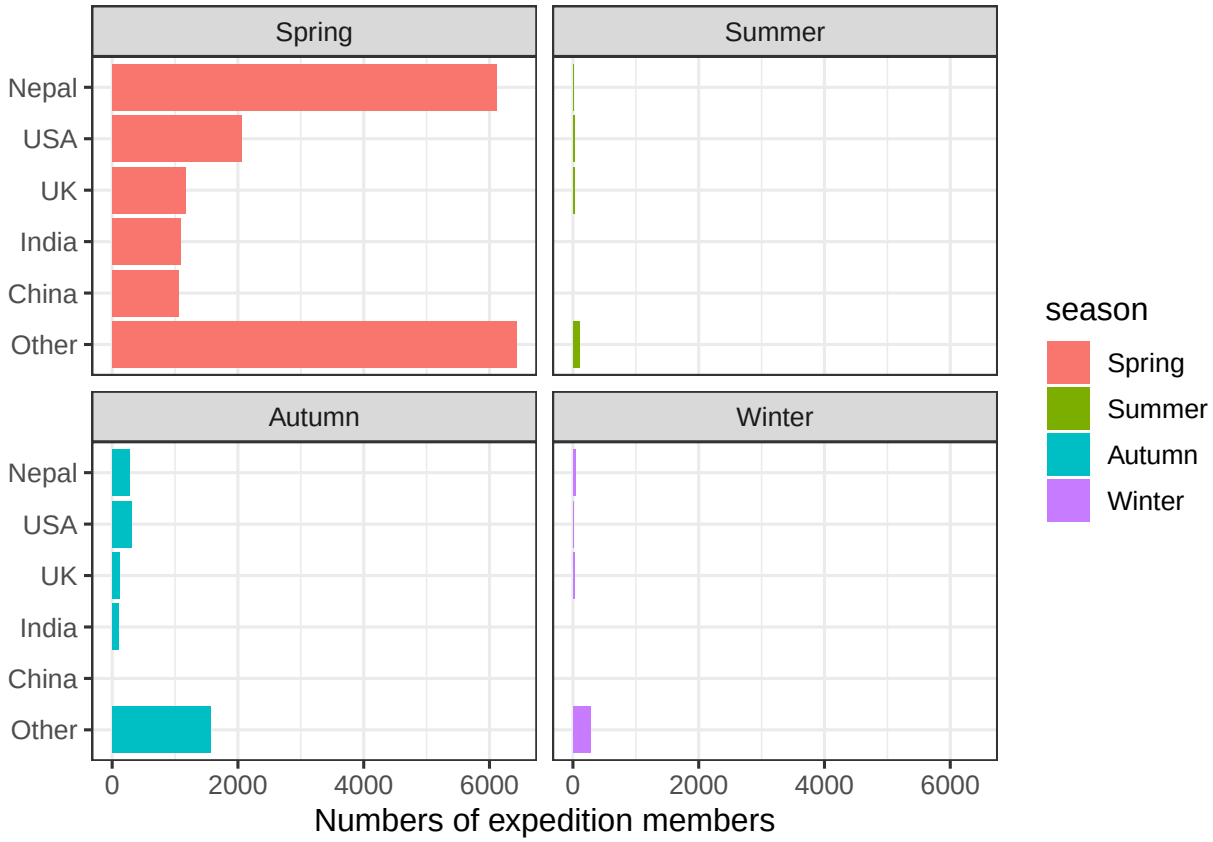
First, I plot the elevation highpoint distribution between the summited group and not summited group using ridgeline plots.

```
ggplot(members_everest, aes(highpoint_metres, sex, fill = success)) +
  geom_density_ridges(scale = 3,
                      rel_min_height = 0.01,
                      alpha = .5,
                      bandwidth = 954,
                      na.rm = TRUE) +
  scale_x_continuous(name = "Elevation highpoint (meter)",
                     limits = c(2500, 12500),
                     expand = c(0, 0)) +
  scale_y_discrete(name = "Sumitted a peak or not",
                   labels = c ("Not summited", "Summited"),
                   expand = c(0, 0)) +
  scale_fill_manual(name = NULL,
                    labels = c("Female", "Male"),
                    values = c('FALSE' = "#56B4E9", 'TRUE' = "#E69F00")) +
  theme(axis.text.y = element_text(vjust = 0))
```



Then I plot the numbers of top 5 countries based on the expedition members' citizenship using bar plots. I facet by season so I can clearly show how the number of expedition members are changed.

```
members_everest$season <- factor(members_everest$season,
                                 levels = c("Spring", "Summer", "Autumn", "Winter"))
members_everest %>%
  mutate(citizenship = fct_lump_n(fct_infreq(citizenship), 5)) %>%
   ggplot(aes(y = fct_rev(citizenship), fill = season)) +
  geom_bar() +
  scale_x_continuous(name = "Numbers of expedition members") +
  scale_y_discrete(name = NULL) +
  facet_wrap(vars(season), ncol = 2) +
   theme_bw(12)
```



Discussion:

For both female and males members of the climbing expedition, the status of summited or not summited a peak does not have much effect on the elevation highpoint. Both female and male groups show similar elevation highpoints regardless of the status of summited or not. Overall, male members show higher elevation highpoints than female members in both summited and not summited groups.

Nepal, USA, UK, India, and China are the 5 countries with the highest numbers of climbing expedition members. The orders of the 5 countries change depending on the seasons. Spring has the highest number of expedition members among the 4 seasons. The second highest number of expedition members is in Autumn. Thus, I can conclude that there is a large number changes of expedition members across the different seasons.