Project 2

In this project, you will be working with a dataset about the members of Himalayan expeditions:

```
members <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/
tidytuesday/master/data/2020/2020-09-22/members.csv')

members
```

```
# A tibble: 76,519 × 21
   expedition_id member_id    peak_id peak_name   year season sex     age
   <chr>         <chr>        <chr>   <chr>      <dbl> <chr>  <chr> <dbl>
 1 AMAD78301     AMAD78301-01 AMAD    Ama Dablam  1978 Autumn M        40
 2 AMAD78301     AMAD78301-02 AMAD    Ama Dablam  1978 Autumn M        41
 3 AMAD78301     AMAD78301-03 AMAD    Ama Dablam  1978 Autumn M        27
 4 AMAD78301     AMAD78301-04 AMAD    Ama Dablam  1978 Autumn M        40
 5 AMAD78301     AMAD78301-05 AMAD    Ama Dablam  1978 Autumn M        34
 6 AMAD78301     AMAD78301-06 AMAD    Ama Dablam  1978 Autumn M        25
 7 AMAD78301     AMAD78301-07 AMAD    Ama Dablam  1978 Autumn M        41
 8 AMAD78301     AMAD78301-08 AMAD    Ama Dablam  1978 Autumn M        29
 9 AMAD79101     AMAD79101-03 AMAD    Ama Dablam  1979 Spring M        35
10 AMAD79101     AMAD79101-04 AMAD    Ama Dablam  1979 Spring M        37
# ℹ 76,509 more rows
# ℹ 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>
```

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/.

Hints:

- Make sure your two questions are actually questions, and not veiled instructions to perform a particular analysis.
- Remember your code needs to contain at least three data manipulation functions for data wrangling before you plot. You are allowed to put all the data wrangling into the answer for one of the two questions.
- You should make one plot per question.
- For at least one plot, you have to use either faceting or color coding or both. Pick whichever you prefer.

- Adjust `fig-width`, `fig-height`, and `out-width` in the chunk options to customize figure sizing and figure aspect ratios. `fig-width` and `fig-height` are given in inches and will usually be between 3 and 10. `out-width` is given in percent and will usually be between 50% and 100%.
- You can use additional R packages such as ggforce, colorspace, etc., if you find them helpful. However, please stick to packages we have discussed in class.

You can delete these instructions from your project. Please also delete text such as *Your approach here* or * *

Question 1: *Across countries is there a relationship between whether a member is male or female and the highpoint that members reach?*

Question 2: *What is the distribution of the success status based on the member's age.*

Introduction: *We are working with the `members` dataset which contains the records for all expeditions that have been climbed in the Nepalian Himalayas from 1905 to Spring 2019. In this dataset, each row in this dataset corresponds to one expedition member, and there are 21 columns providing information about the expedition and member details partaking in the expedition. Information about the expedition includes expedition id, peak id, peak name, year, and season. Information about member details partaking in the expedition include sex, age, citizenship, expedition role, hired status, highpoint of the member, success status, solo status, oxygen used, death status, death cause, death height, injured status, injury type, and height of injury. To answer the two questions, we will work with three variables, the member's sex(column sex), the members citizenship(column citizenship), the members highpoint elevation(column highpoint_metres), whether the expedition of the member was successful or not(column success), and the members age(column age). The members sex is provided a categorical value as M or F, the members citizenship is provided as a categorical value as the name of the country the member is from, the members highpoint elevation is provided a numeric value in meters, the success status is provided as categorical value as either TRUE or FALSE, and the member's age is provided as a numeric value in years. *

Approach: *To show the relationship of the highpoint elevation of a member per member's country and sex we will be using a scatterplot (geom_point()). We seperated the highpoint, sex, citizenship and only included the twenty most frequent country(to make the visualization legible) and used the average highpoint elevation to do visualization/relationship. Scatterplots is the best way to display relationships with at least one numeric variable with categorical variables also included. To show the distributions of the success status depending on the member's age we will be using violin plots (geom_violin()).Violins are the best way to show the distributions of a categorical variable and in this case the x value will be the success rate as to best visualize it.*

Analysis: Question 1: *Across countries is there a relationship between whether a member is male or female and the highpoint that members reach?* **To answer this question**

```
# Q1: Your R code here

library(cowplot)
```

```
Attaching package: 'cowplot'
```

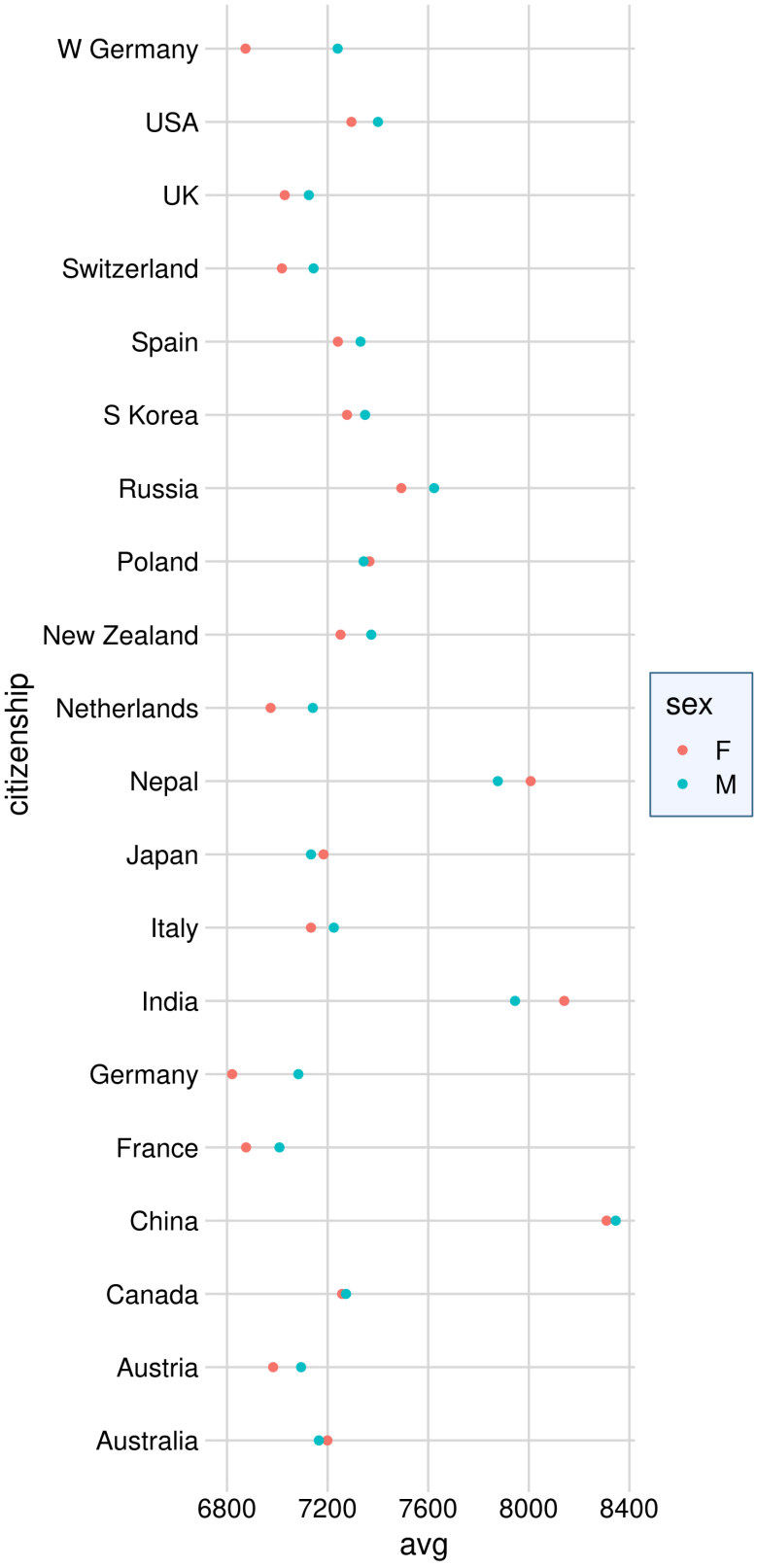
```
The following object is masked from 'package:lubridate':

    stamp
```

```
country <- members |>
  mutate(citizenship = fct_infreq(citizenship)) |>
  count(citizenship)  |>
  slice(1:20) |>
  as.list(members[citizenship])


members |>
  select(citizenship, highpoint_metres, sex) |>
  filter(citizenship %in% c(country$citizenship))|>
  drop_na() |>
  group_by(citizenship, sex) |>
  summarize(n = n(), avg = mean(highpoint_metres)) |>
  ggplot() +
  aes(
    avg, citizenship
  )  +
  geom_point(aes(color = sex)) +
  theme_minimal_grid() +
  theme(
    legend.box.background = element_rect(
      fill = "aliceblue",
      color = "steelblue4" # line color
    ),
    legend.box.margin = margin(7, 7, 7, 7)
  )
```

```
`summarise()` has grouped output by 'citizenship'. You can override using the
`.groups` argument.
```


W Germany
USA
UK
Switzerland
Spain
S Korea
Russia
Poland
New Zealand
Netherlands
Nepal
Japan
Italy
India
Germany
France
China
Canada
Austria
Australia
citizenship
6800
7200
7600
8000
8400
avg
sex
F
M

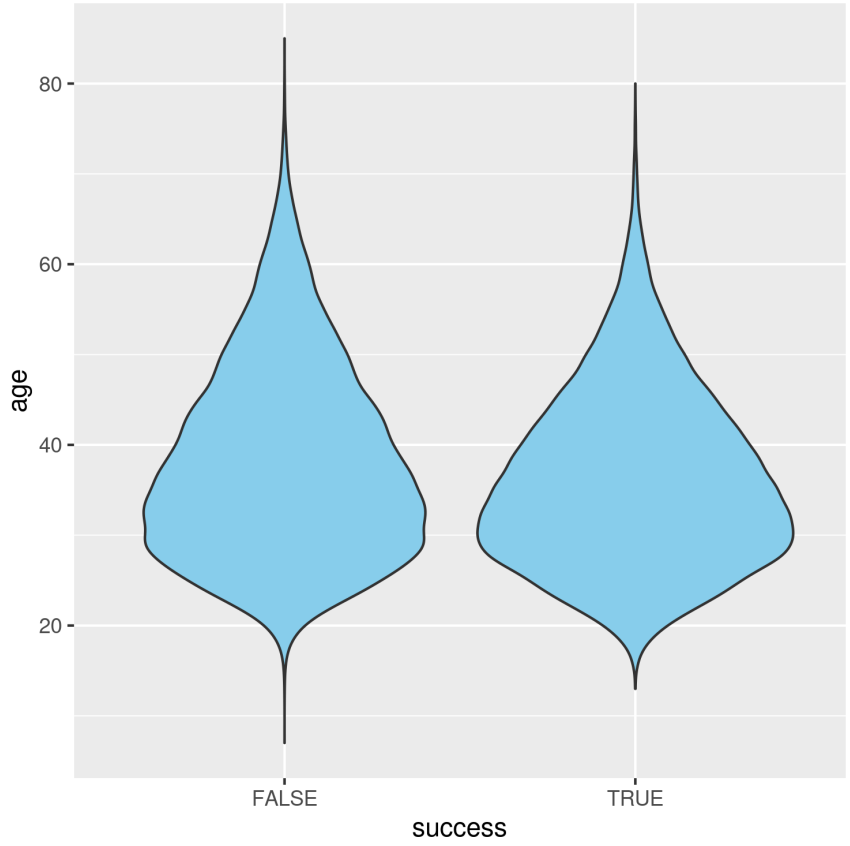
W Germany and Germany have the greatest difference between high point elevation between the sexes of about 200-400 meters.

Question 2: *What is the distribution of the success status based on the member's age.* **To answer this question**

```
# Q2: Your R code here

ggplot(members, aes(x = success, y = age)) +
  geom_violin(fill = "skyblue")
```

```
Warning: Removed 3497 rows containing non-finite outside the scale range
(`stat_ydensity()`).
```



The age of those who succeeded and failed were around in their 30's Discussion: Sex of a member has a slight difference in the highpoint elevation that one does reach. Although for countries like W Germany and Germnay there is a massive difference in highpoint elevation of about 200-400 meter with the males reaching higher elevation. We can see this in the fact there is a slight gap in between the point in the scatterplot although it is not terribly large. Their doesn't seems to be a relationship between both sex and country a member is from on the highpoint elevation

but relationship between sex per country in highpoint elevation with male members for the top twenty countries having higher elevation than female members. A possible explanation for some countries having more a discrepancy betwen sexes could be due to the fact that more men signed up for the expeditions that women especially in the early days because this data was taken over many years. Survey analysis for the sign ups of these expeditions over time would be needed in a line graph to determine if that is the case. For the success rate per age while the bulk of people who both failed and succeded in their expedition is around the same age of being in their 30's their distributions differed. Their is less older people have succeed in the expedition that younger people and more older people who have failed. Also younger people who have succeed as well with their being clusters of people under their 20's who have succeeded. A possible explanation for all of this is that younger people are more fit that older people on average while the average person who can climb these expedition are in their 30's which accounts for the average ages for those who succeed or not. Also another factor could be that not a lot of older people don't go on these expeditions in the first place as advised by doctors and family members. Further analysis would be to gather these sociological dat points to provided an explanation for said results we have found. **