Project 2

Brian Folkers (bdf676)

This is the dataset you will be working with:

```
members <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/20
```

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: Looking only at expeditions to Mt.Everest since 1960, how do deaths in each season break down by the seven most common causes?

To answer this question, create a summary table and one visualization. The summary table should have 4 columns: "death_cause", "Spring", "Summer", "Autumn" and "Winter", where the seasons columns have the raw number of deaths for each cause in the first column. Remember to replace any `NA` values with `0`.

We recommend you use faceted pie charts for the visualization. The visualization should show the relative proportion of the 7 most common death causes for each season. Include an additional category called "other" for all other death causes.

Please note that we are not asking you to find the seven most common causes of death separately for each season. Find the seven most common causes of death overall and then perform the analysis by season.

Introduction: The `members` dataset contains data on all expeditions to the Himalayan Mountains, including data on deaths that occurred on the expedition. In this part of the project, the goal will be to determine what the breakdown for causes of death are, by season, for expeditions to Mt.Everest since 1960. To answer this, this section will largely focus on the `death_cause` and `season` variables (it will also be limited by `year` and `peak_name`).

Approach: The approach for this section will be broken into two parts: a summary table of deaths by season, and a faceted pie chart visualization.

For the summary table, the data wrangling will focus on:

- limit the data to expeditions since 1960 & only Mt. Everest
- create a count of deaths
- split `season` into four new columns: `Spring`, `Summer`, `Autumn` and `Winter`
- create a `Total` deaths column and order it by descending value

For the visualization, the data wrangling will focus on:

- group together all `death_cause` values into other if they are not in the top 7 causes

The faceted pie chart will be ideal for this visualization as it will allow for a better understanding of the proportion of death causes for each season.

Analysis:

```
#data wrangling for summary table
members_new <- members %>% #new dataframe
  filter(year >= 1960, died == "TRUE", peak_name == "Everest") %>% #since 1960 & only Everest
```

```r
  count(death_cause, season)  %>% #makes a count of deaths
  filter(season != "Unknown") %>% #filters out unknown season
  pivot_wider(names_from = "season", values_from = "n") %>% #splits up seasons into seperate columns
  mutate(Spring = ifelse(is.na(Spring), 0, Spring),
         Summer = ifelse(is.na(Summer), 0, Summer),
         Autumn = ifelse(is.na(Autumn), 0, Autumn),
         Winter = ifelse(is.na(Winter), 0, Winter)) %>% #replaces seasons NAs with 0s
  filter(!is.na(death_cause)) %>% #removes NA category from death_cause column
  mutate(Total = Autumn + Spring + Winter + Summer) %>% #makes total deaths column
  arrange(desc(Total)) #sorts by total deaths largest to smallest

#view data frame
members_new
```

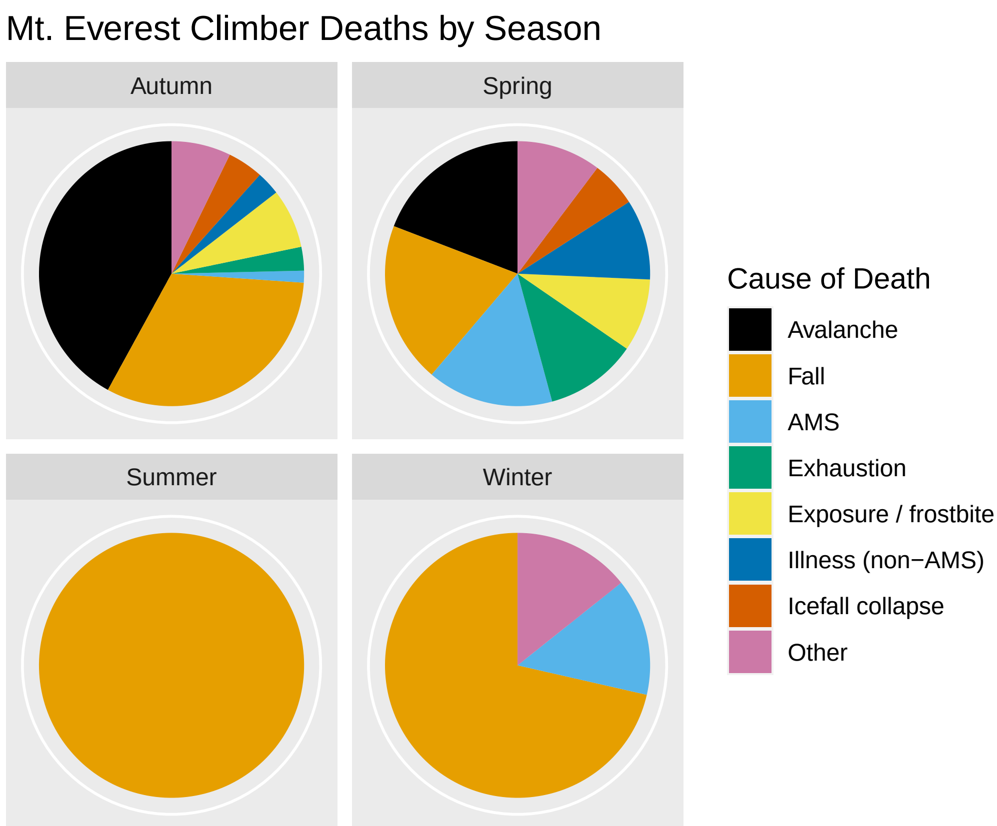
```
## # A tibble: 12 x 6
##    death_cause                 Autumn Spring Winter Summer Total
##    <chr>                        <dbl>  <int>  <dbl>  <dbl> <dbl>
##  1 Avalanche                       29     41      0      0    70
##  2 Fall                            22     42      5      1    70
##  3 AMS                              1     33      1      0    35
##  4 Exhaustion                       2     24      0      0    26
##  5 Exposure / frostbite             5     19      0      0    24
##  6 Illness (non-AMS)                2     21      0      0    23
##  7 Icefall collapse                 3     12      0      0    15
##  8 Crevasse                         2      8      1      0    11
##  9 Disappearance (unexplained)      0      8      0      0     8
## 10 Other                            3      2      0      0     5
## 11 Falling rock / ice               0      2      0      0     2
## 12 Unknown                          0      2      0      0     2
```

```r
#data wrangling for visualization
members_viz <- members %>% #new dataframe
  filter(year >= 1960, died == "TRUE", peak_name == "Everest") %>% #since 1960 & Everest
  mutate(death_cause = fct_lump_n(fct_infreq(death_cause), 7, other_level = "Other")) %>% #groups small
  count(death_cause, season)  %>% #makes a count of deaths
  filter(season != "Unknown") %>% #filters out unknown season
  filter(!is.na(death_cause)) %>% #removes NA category from death_cause column
  group_by(season)

#view visualization dataframe
members_viz
```

```
## # A tibble: 20 x 3
## # Groups:   season [4]
##    death_cause          season     n
##    <fct>                <chr>  <int>
##  1 Avalanche            Autumn    29
##  2 Avalanche            Spring    41
##  3 Fall                 Autumn    22
##  4 Fall                 Spring    42
##  5 Fall                 Summer     1
##  6 Fall                 Winter     5
##  7 AMS                  Autumn     1
##  8 AMS                  Spring    33
##  9 AMS                  Winter     1
```

```
## 10 Exhaustion           Autumn     2
## 11 Exhaustion           Spring    24
## 12 Exposure / frostbite Autumn     5
## 13 Exposure / frostbite Spring    19
## 14 Illness (non-AMS)    Autumn     2
## 15 Illness (non-AMS)    Spring    21
## 16 Icefall collapse     Autumn     3
## 17 Icefall collapse     Spring    12
## 18 Other                Autumn     5
## 19 Other                Spring    22
## 20 Other                Winter     1
```

```r
#package for Okabe-Ito theme
library(ggthemes)

#visualization: ggplot faceted pie chart
ggplot(members_viz) +
  aes(n, "YY", fill = death_cause) + # death_cause fill & end_angle for units
  geom_col(position = "fill") +
  ggtitle("Mt. Everest Climber Deaths by Season") +
  labs(fill = "Cause of Death") +
  facet_wrap(vars(season)) + #facet by season
  coord_polar() +
  scale_fill_colorblind() + #uses Okabe-Ito colors
  scale_x_continuous(
    name = NULL, breaks = NULL
  ) +
  scale_y_discrete(
    name = NULL, breaks = NULL
  )
```



Discussion: For the summary table, there are a couple trends that stand out. Firstly, `Spring` and `Autumn` seem to have significantly higher death counts both in total and in the same categories as `Winter` and `Summer`. Second, deaths by `Avalanche` and `Fall` seem the most common by far. For the visualization, a somewhat similar trend can be noticed in regards to the seasons. `Winter` and `Summer` have relatively few causes of death (mostly `Fall`). This may be due to these seasons having relatively fewer deaths as previously mentioned. `Autumn` has more variation in the causes of deaths, with most still being `Avalanche` and `Fall`. Finally, `Spring` has the most variation in causes of death, with no cause having a clear majority. This analysis has shown that expeditions to Everest are the deadliest in `Spring` and `Autumn`, both in raw counts as well as the causes of death.

Part 2

Question: Looking at all expeditions, is there a relationship between the `sex`, `age`, and `citizenship` of the climbers and the `success` of the expedition?

Introduction: In this part of the project, the goal will be to determine if the different demographic variables like age distribution, sex or and the country of origin of the climbers differs between successful and unsuccessful expeditions. To answer this, this section will largely focus on the `age`, `citizenship`, `sex` and `success` variables. The data contains a wide variation of both ages and countries, so it is likely that these variables will have some relationship with the success of the expeditions.

Approach: The data wrangling for this section focused on a few simple modifications. For the summary table, the count of ages for the top 5 most frequent countries was made, similar to part 1. For the visualization data, NAs for all needed variables were removed, and the least frequent appearing countries outside of the top 10 were placed into the Other category. For the visualization a faceted density plot will allow for the analysis of both the age distribution and how it relates to sex, success of the expedition, and the country of origin.

Analysis:

```r
#data wrangling for summary table
members_new2 <- members %>% #new dataframe
  mutate(citizenship = fct_lump_n(fct_infreq(citizenship), 5, other_level = "Other")) %>% #lumping toge
  count(citizenship, age)  %>% #makes a count of age
  filter(!is.na(age)) %>% filter(!is.na(citizenship)) %>% #removing NAs
  pivot_wider(names_from = "citizenship", values_from = "n") %>% #splits up citizenship into seperate c
  mutate(Nepal = ifelse(is.na(Nepal), 0, Nepal),
         USA = ifelse(is.na(USA), 0, USA),
         Japan = ifelse(is.na(Japan), 0, Japan),
         UK = ifelse(is.na(UK), 0, UK),
         France = ifelse(is.na(France), 0, France),
         Other = ifelse(is.na(Other), 0, Other))

#view data frame
members_new2
```

```
## # A tibble: 74 x 7
##      age Nepal   USA Japan    UK France Other
##    <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>
##  1    12     1     0     0     0      0     0
##  2    14     1     0     0     0      1     2
##  3    15     3     2     0     0      1     4
##  4    16     7     5     0     2      3    22
##  5    17    16     7     0     1      6    36
##  6    18    53     6     4    10      7    71
##  7    19   118     8    22    20     12   102
##  8    20   198    24    37    15      8   166
##  9    21   281    15   118    22     22   263
## 10    22   364    35   146    41     34   429
## # ... with 64 more rows
```

```r
#data wrangling for visualization
members_viz2 <- members %>% #new dataframe
  select(citizenship, age, success, sex) %>% #selecting needed varaibles
  mutate(citizenship = fct_lump_n(fct_infreq(citizenship), 10, other_level = "Other")) %>% #lumping tog
  filter(!is.na(age)) %>% filter(!is.na(citizenship)) %>% filter(!is.na(success)) %>% #removing NAs
  mutate(success = replace(success, success == FALSE, "Failure")) %>% #renaming
  mutate(success = replace(success, success == TRUE, "Success")) %>% #renaming
  mutate(sex = replace(sex, sex == "M", "Male")) %>% #renaming
  mutate(sex = replace(sex, sex == "F", "Female")) #renaming

#view visualization dataframe
members_viz2
```

```
## # A tibble: 73,021 x 4
##    citizenship   age success sex
##    <fct>       <dbl> <chr>   <chr>
##  1 France         40 Failure Male
##  2 France         41 Failure Male
##  3 France         27 Failure Male
##  4 France         40 Failure Male
##  5 France         34 Failure Male
##  6 France         25 Failure Male
##  7 France         41 Failure Male
```

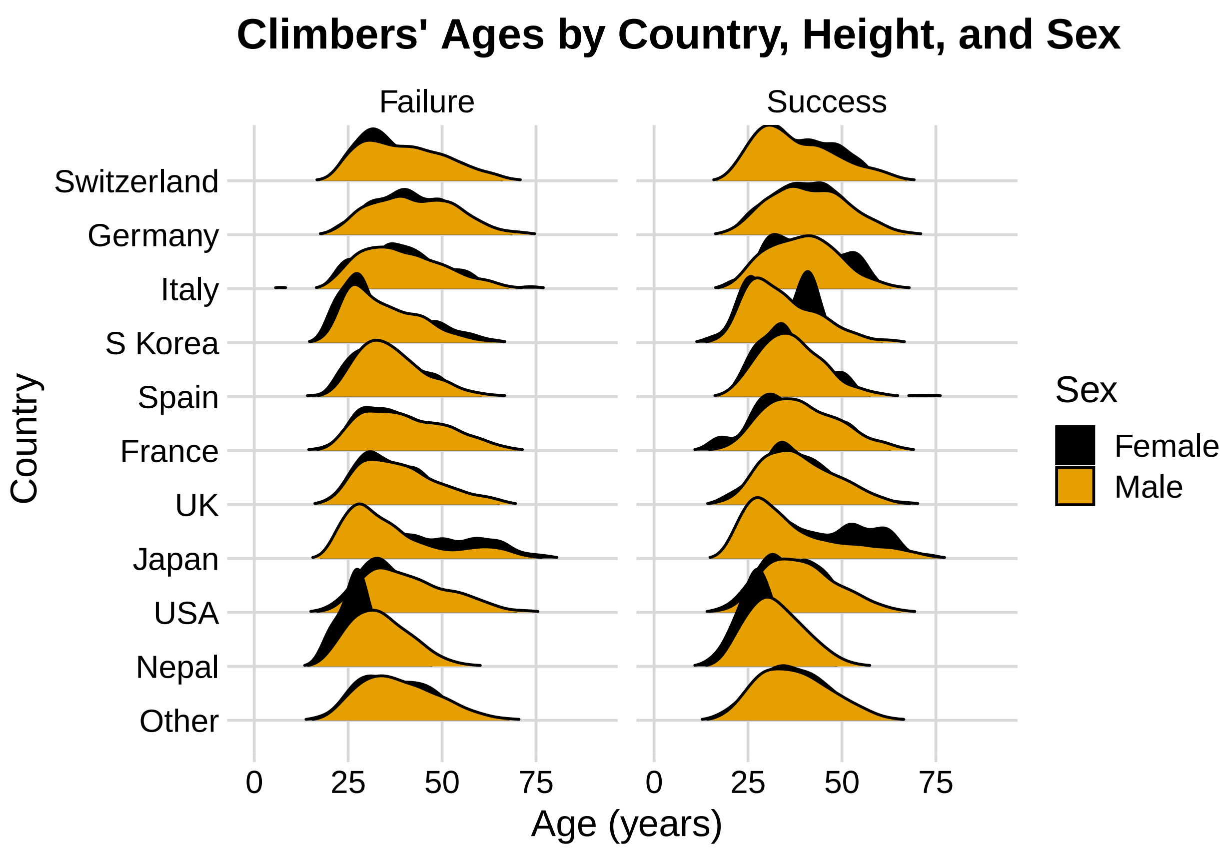
```
##   8 France          29 Failure Male
##   9 USA             35 Failure Male
## 10 Other           37 Success Male
## # ... with 73,011 more rows
```

```r
library(ggridges) #package
library(cowplot) #package
```

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

## The following object is masked from 'package:ggthemes':
##
##     theme_map
```

```r
ggplot(members_viz2, aes(x = age, y = fct_relevel(citizenship, "Other"), fill = sex)) + #reorders to pu
  labs(fill = "Sex") +
  xlab("Age (years)") +
  ylab("Country") +
  ggtitle("Climbers' Ages by Country, Height, and Sex") +
  geom_density_ridges(rel_min_height = 0.01) +
  scale_fill_colorblind() + #uses Okabe-Ito colors
  facet_wrap(vars(success)) + #facet by success
  theme_minimal_grid()
```

```
## Picking joint bandwidth of 2.38

## Picking joint bandwidth of 2.61
```



Discussion: Overall, there does not appear to be a large amount of variation in sex or age for successful and unsuccessful expeditions, but there are a few observations that stand out. From the summary table and the visualization, it is clear that there is a wide range in ages, with a heavy concentration for most countries between 25 and 50. Japan, Italy, and South Korea have a large number of older females with successful expeditions. Nepal has a high concentration of expeditions at comparatively younger ages. Most countries appear to preform relatively similarly in the expeditions. Overall, it would appear that these variables are largely reflections of the demographics of these countries, and do not have a significant outcome on the success of the expeditions.