Project 2

Ethan Chang - ehc586

This is the dataset you will be working with:

```
members <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2

members
```

```
## # A tibble: 76,519 x 21
##    expedition~1 membe~2 peak_id peak_~3  year season sex     age citiz~4 exped~5
##    <chr>        <chr>   <chr>   <chr>   <dbl> <chr>  <chr> <dbl> <chr>   <chr>
##  1 AMAD78301    AMAD78~ AMAD    Ama Da~  1978 Autumn M        40 France  Leader
##  2 AMAD78301    AMAD78~ AMAD    Ama Da~  1978 Autumn M        41 France  Deputy~
##  3 AMAD78301    AMAD78~ AMAD    Ama Da~  1978 Autumn M        27 France  Climber
##  4 AMAD78301    AMAD78~ AMAD    Ama Da~  1978 Autumn M        40 France  Exp Do~
##  5 AMAD78301    AMAD78~ AMAD    Ama Da~  1978 Autumn M        34 France  Climber
##  6 AMAD78301    AMAD78~ AMAD    Ama Da~  1978 Autumn M        25 France  Climber
##  7 AMAD78301    AMAD78~ AMAD    Ama Da~  1978 Autumn M        41 France  Climber
##  8 AMAD78301    AMAD78~ AMAD    Ama Da~  1978 Autumn M        29 France  Climber
##  9 AMAD79101    AMAD79~ AMAD    Ama Da~  1979 Spring M        35 USA     Climber
## 10 AMAD79101    AMAD79~ AMAD    Ama Da~  1979 Spring M        37 W Germ~ Climber
## # ... with 76,509 more rows, 11 more variables: 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>, and abbreviated variable
## #   names 1: expedition_id, 2: member_id, 3: peak_name, 4: citizenship,
## #   5: expedition_role
```

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/.

Question 1: How do the causes of death vary among different seasons?

Question 2: How has the mortality rate changed over the years?

Introduction: For this project, we work with the Himalayan Expeditions (or `members`) dataset, a dataset consisting of information for all expeditions that have climbed in the Nepal Himalayas from 1905 to 2019. Each row in the dataset corresponds to a member from a specific expedition with multiple columns denoting that member's demographic, status, expedition date, and expedition setting. More specifically, the dataset contains a member's id, expedition id, sex, age, peak id, name of peak climbed, year and season climbed, citizenship, expedition role, whether they were hired, succeeded, went solo, used oxygen, died, or injured, their highpoint, how they died, how high they fell, and how they were injured.

To answer the two questions above, we work with 4 main variables from the dataset, the `season` and `year` of the expedition, whether the members `died`, and the cause of death (`death_cause`). The season is given as a character being either Autumn, Spring, Summer, or Winter, while the year is given as numeric value. The death status is presented as a boolean, with TRUE meaning they died and FALSE meaning they did not die. The cause of death is given as a character categorizing how the expedition member died.

Approach: To answer the first question, we would wrangle the data by first filtering (`filter()`) it to those that died and selecting (`select()`) the `season` and cause of death variables as those are the only two we are

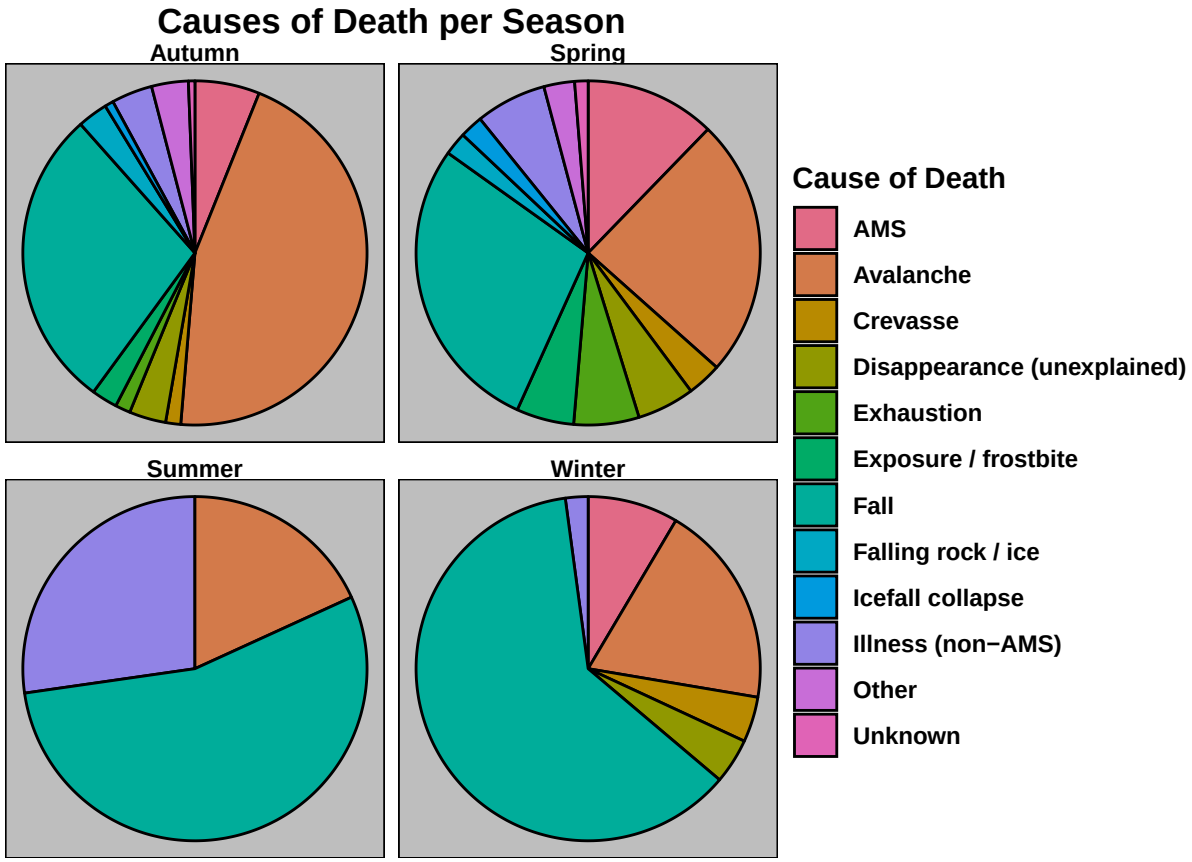
interested in. We would then group by (`group_by()`) those two variables and count the total number of each `season` and cause of death combination. To show the proportions of causes of deaths between season, we used pie charts (`geom_arc_bar(stat = "pie")`) as this method highlights how each part is a fraction of a whole, in this case the multitude of causes of death. It emphasizes which part(s) are more prevalent in a more concise and compact manner, facilitating comparisons between multiple parts over multiple graphs. This was important as we would then facet it by `season` in order to compare the pie charts between the four seasons.

To answer the second question, we would wrangle the data by first mutating (`mutate()`) a `decade` variable by cutting (`cut()`) the `year` variable into bins of 10 from 1900 to 2020. We do this as simply looking at each individual year is too narrow and less visually appealing; grouping it solves this problem and makes it easier to compare between segments. We would then group by `decade` and summarize (`summarize()`) the data set into two new variables: the total number of deaths (`num_died`) and percentage of deaths (`percent_died`). To show the amount of deaths over the years, which we binned into decades, we used barplots (`geom_col()`) as it facilitates side by side comparisons of amounts using height over the decades. We would then color map each bar, or `decade`, by the mortality rate in order to highlight what proportion of expedition members actually died in that specific decade for comparison and analysis purposes.

Analysis:

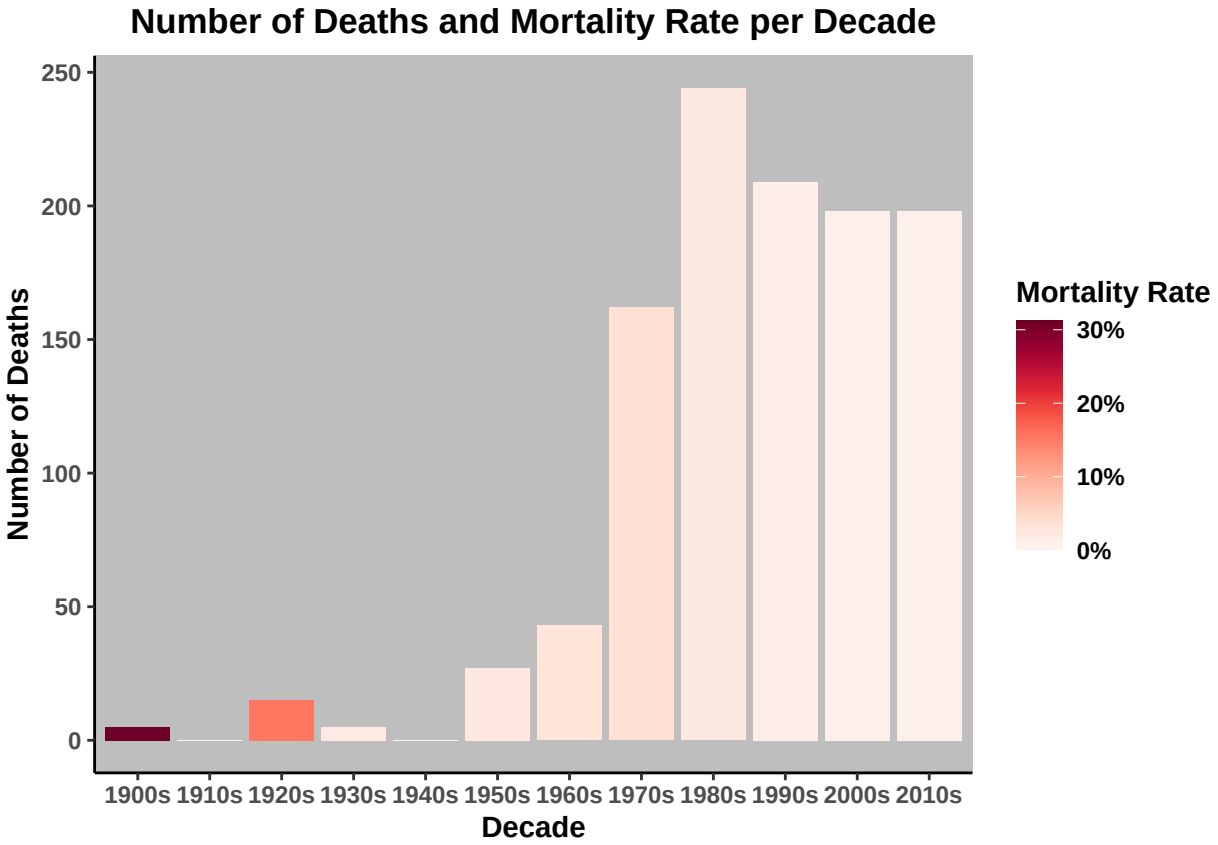
```r
# data wrangling for death causes grouped by season
deathcause_per_season <- members %>%
  filter(died == TRUE) %>%
  select(c(season, death_cause)) %>%
  group_by(season, death_cause) %>%
  summarize(count = n(), .groups = "drop_last")

# plot pie charts showing proportion of death causes faceted per season
ggplot(deathcause_per_season) +
  aes(x0 = 0, y0 = 0, r0 = 0, r = 1, amount = count, fill = death_cause) +
  geom_arc_bar(stat = "pie") +
  coord_fixed() +
  facet_wrap(~season) +
  labs(title = "Causes of Death per Season", fill = "Cause of Death") +
  theme_void() +
  scale_fill_discrete_qualitative(palette = "Dark 3") +
  theme(text = element_text(face = "bold"),
        plot.title = element_text(hjust = 0.5),
        panel.background = element_rect(fill = "gray"))
```




```
# data wrangling for number and percentage of deaths per decade
mortality_decade <- members %>%
  # bins years into decades
  mutate(decade = cut(year, breaks = seq(1900, 2020, 10), dig.lab = 4)) %>%
  group_by(decade) %>%
  summarize(num_died = sum(died), percent_died = sum(died)/n() * 100)

# plot bar graphs showing number of deaths per decade, color mapped by percentage of deaths
ggplot(mortality_decade, aes(decade, num_died, fill = percent_died)) +
  geom_col() +
  scale_x_discrete(name = "Decade", labels = paste0(seq(1900,2010,10), "s")) +
  scale_y_continuous(name = "Number of Deaths") +
  ggtitle("Number of Deaths and Mortality Rate per Decade") +
  theme_classic() +
  scale_fill_continuous_sequential(name = "Mortality Rate",
                                   labels = paste0(seq(0,30,10), "%"),
                                   palette = "Reds") +
  theme(text = element_text(face = "bold"),
        plot.title = element_text(hjust = 0.5),
        panel.background = element_rect(fill = "gray"))
```



Discussion: Based on the first figure produced in our analysis, it can be seen that between all seasons, the most prevalent causes of death (at least 50%) seem to come from falls and avalanches. During the summer, non-AMS illnesses seem to also be rather prevalent, while other seasons tend to see more AMS and small proportions of the other causes. These distributions make sense as it is expected that expedition members are fully trained to tackle the harsh climate and environment of the peaks they are travelling through, so if they were to die, it would most likely be from accidents or uncontrollable circumstances like falls, avalanches, and illnesses. During the summer and winter, falls account for over 50% of deaths, likely due to the extreme temperatures and weather conditions that may make accidental slips and falls more possible. During the autumn, avalanches seem to be more prevalent, likely because it follows the higher summer temperatures that may have affected the structural integrity of the snow; while spring sees a more even distribution between falls and avalanches, likely due to it warming up a little after the harsh winter climate.

From our second figure, we can see that over the decades, there has been an increasing trend in the total number of deaths, but a decreasing trend in the mortality rate, or percentage of deaths. This implies that while the total number of expedition members dying is increasing, so is the total number of expedition members going on expeditions, at a larger rate. This makes sense since if more members are going on expeditions, more people are likely to die in the process, as there would be more chances for it to happen. The lowering mortality rate is also expected as over the years, knowledge, strategies, technological advancements, training, and safety equipment have only been improving, meaning that it is less likely for expedition members to die.