Project 2

Enter your name and EID here

Connor Hanna cdh3663

This is the dataset you will be working with:

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

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:

Data comes from The Himalayan Database, containing the information concerning a specific expedition in the Nepalese portion of the Himalayas. At the observation level the data includes information about each member of the expedition, identified by the `member_id` variable. The data includes information about expeditions from 1905 to 2019 for more than 490 peaks in Nepal, and is encoded in a long format. Variables include information concerning oxygen usage, age of the participant, the year and season of the expedition, the name of the peak, if any injury occurred, and so on.

To answer the question, I'll need information from the variables `year`, `season`, `peak_name`, `died`, and `death_cause`. `year` is a numeric variable storing the year the expedition took place. `season` is a categorical variable describing the season of the expedition as autumn, summer, spring, or winter. `peak_name` is a categorical variable containing the peak the expedition took place on, and `died` is a dummy variable describing if the individual died during the expedition. `death_cause` is a categorical variable that represents the cause of death in a character string.

Approach:

For the summary table I'll first use `filter()` and `select()` to pare down the data to only the information on season and cause of death for Everest attempts after 1960. Then, I'll use `table()` and `as.data.frame.matrix()` to generate an appropriate summary table.

For the pie chart portion, I'll begin by doing some additional data cleaning. First I'll use `count()` , `arrange` , `head()` and `as.list()` to store the top seven causes of death. I then define the function `%notin%` and use

it to replace the causes of death that weren't in the top seven with "Other". Afterwards, `group_by()` and `summarize()` will then condense the data into a table with clearly defined arguments for `ggplot()`. Finally, I'll use `ggplot()` and `geom_arc_bar(stat = "pie")` to generate the faceted pie charts.

Faceted pie charts were chosen because they efficiently describe proportions. Since cause of death varies by season, faceting is necessary for comparison.

The cross table was chosen to facilitate an understanding of the variables used in construction of the visualization.

Analysis:

```
#filtering for deaths on Everest after 1960
members <-
  members_og |>
    filter(peak_name == "Everest") |>
    filter(year >= 1960) |>
    filter(died == TRUE)

#selecting the relevant variables for the table
#we're already done filtering so peak name is no longer needed
members <-
  members |>
    select(season, death_cause)
```

```
#great! Now for the table
table <-
  table(members$death_cause, members$season)

#that result was great, but it wasn't saved as a dataframe.
summary_table <-
  as.data.frame.matrix(table)
summary_table
```

```
##                               Autumn Spring Summer Winter
## AMS                                1     33      0      1
## Avalanche                         29     41      0      0
## Crevasse                           2      8      0      1
## Disappearance (unexplained)        0      8      0      0
## Exhaustion                         2     24      0      0
## Exposure / frostbite               5     19      0      0
## Fall                              22     42      1      5
## Falling rock / ice                 0      2      0      0
## Icefall collapse                   3     12      0      0
## Illness (non-AMS)                  2     21      0      0
## Other                              3      2      0      0
## Unknown                            0      2      0      0
```

```
#thanks to the table we just generated, we have most of the information we need all in one place.
#but we still need the seven most common causes of death
cause_count <-
  members |>
    count(death_cause) |>
    arrange(desc(n))
cause_count
```

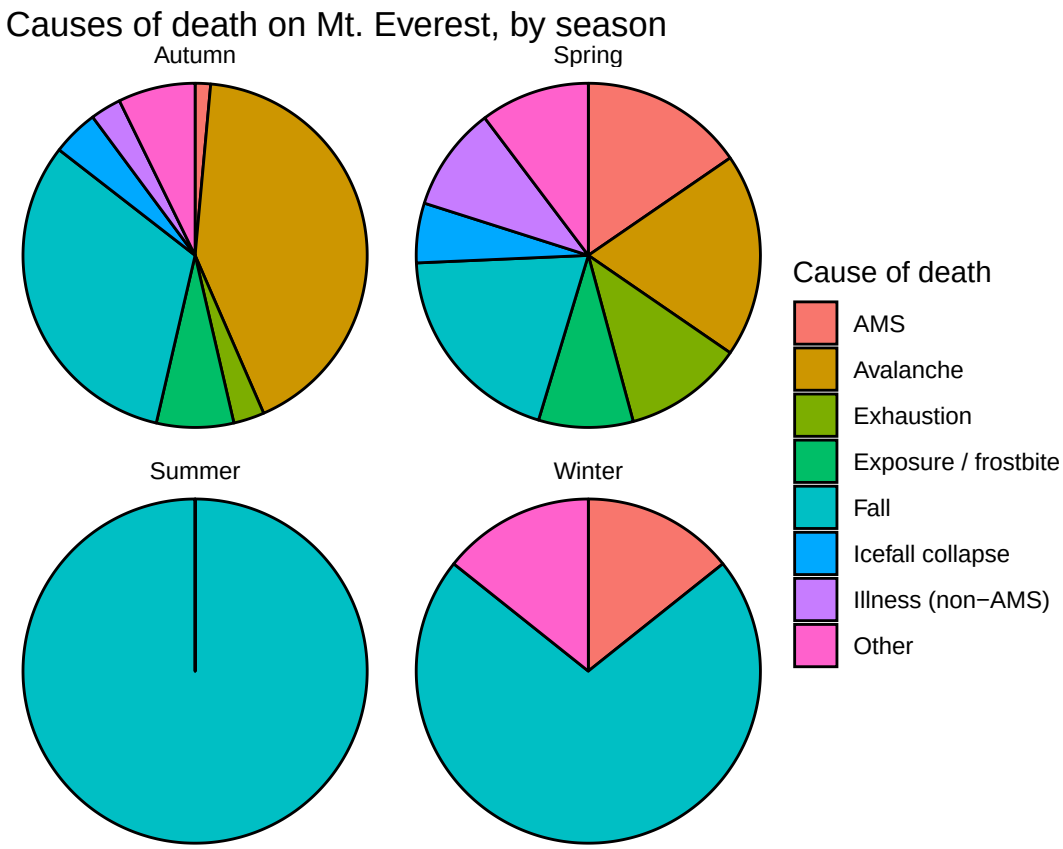
```
## # A tibble: 12 x 2
##    death_cause                    n
##    <chr>                      <int>
##  1 Avalanche                     70
##  2 Fall                          70
##  3 AMS                           35
##  4 Exhaustion                    26
##  5 Exposure / frostbite          24
##  6 Illness (non-AMS)             23
##  7 Icefall collapse              15
##  8 Crevasse                      11
##  9 Disappearance (unexplained)    8
## 10 Other                          5
## 11 Falling rock / ice             2
## 12 Unknown                        2
```

```
#now we have the top causes of death, so we can make a list
top7 <-
  cause_count |>
    head(n = 7)

top7 <-
  as.list(top7$death_cause)

#then we use the list to replace values in the death_cause column with "other"
`%notin%` <- negate(`%in%`)

condition <- members$death_cause %notin% top7

members$death_cause[condition] <- "Other"

#using group_by() and summarize() to generate a table with clear aesthetic assignments for ggplot

members <-
  members |>
    group_by(death_cause, season) |>
    summarize(
      n = n()
    )
```

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

```
#and now to generate our pie chart
members |>
  ggplot(
    aes(
      x0 = 0, y0 = 0,
      r0 = 0, r = 1,
      amount = n,
      fill = death_cause,
    )
  ) +
      geom_arc_bar(stat = "pie") +
```

```
    theme_void() +
    coord_fixed() +
    labs(title = "Causes of death on Mt. Everest, by season", fill = "Cause of death") +
    facet_wrap(vars(season))
```



Discussion:

Looking at the pie charts, fatalities during summer and winter are overwhelmingly from falls. During the fall and spring the increase in deaths from avalanches, icefalls, and exposure imply that weather changes lead to unstable snow packs and treacherous conditions on the mountain. Deaths from exhaustion, altitude sickness, and illness also increased in the transitional seasons - but this mortality isn't easily explained by weather.

Looking at the cross table generated in the second code block, it's clear that the overwhelming majority of fatalities during Everest expeditions occur in the spring and fall. If the pie charts had attributed additional deaths primarily to weather-related causes, it would be theoretically possible for seasonal changes to be a primary explanatory variable with traffic remaining constant. Unfortunately for that explanation, nearly half of Spring and a quarter of Fall deaths are attributed to non-climate causes.

Part 2

Question:

What is the seasonal distribution of expeditions to Mt. Everest? Further, do total deaths in a season correlate closely with increases in expeditions?

Introduction:

The visualizations and data exploration from Part 1 instigate more questions than they provide answers. Why do summer and winter have so few deaths? Why is the death composition of autumn different from spring? We stripped the non-death observations from the data in the first code block, but I have a feeling that the mountain is unassailable in the summer and winter and that a seasonal concentration of traffic explains the pattern in both concentration and types of deaths. I also noticed that the deaths increased over time - was this due to an increase in traffic on the mountain, or is Everest becoming more dangerous?

To answer the question, I'll need information from the variables `year`, `season`, `peak_name`, `died`, `death_cause`, and `expedition_id`.

`expedition_id` is a categorical variable assigning the individual from the observation to a specific expedition. `year` is a numeric variable storing the year the expedition took place. `season` is a categorical variable describing the season of the expedition as autumn, summer, spring, or winter. `peak_name` is a categorical variable containing the peak the expedition took place on, and `died` is a dummy variable describing if the individual died during the expedition. `death_cause` is a categorical variable that represents the cause of death in a character string.

Approach:

First I'll use `select()` and `filter()` to strip the data down to observations after 1960, with the variables `year`, `season`, `died`, `death_cause`, and `expedition_id`. Then I will use `distinct()` to remove duplicate entries for each expedition.

I will use the resulting dataframe to generate a cross table of expeditions by year and season. I chose these variables for the table since they succinctly answer the big questions I was left with after reviewing the table from Part 1: what is the distribution of expeditions across time?

After generating my table, I will use `count()` to collapse the data from individual expeditions into a tally for each year and season combination.

From there it will be fairly straightforward to generate a grouped line graph for the annual expedition count of each season. This visualization captures two important features I wanted to highlight - the proportional variation in Autumn and Spring expeditions, and the massive increase in expeditions beginning around the year 1995. It conveniently allows me to simultaneously explore two continuous variables and one discrete one, where a pie chart would be limited two one continuous variable and one or two categorical. I thought about choosing tiles, but felt that this conveyed proportion better.

In order to generate the death statistics for my final graph, I will first use `filter()` and `select()` to generate a separate dataframe containing incidences of death by season and year. I will then use `distinct()` and `count()` to condense the data from observations of individual deaths into a list of tallies by season and year. I will then `rename()` the variables containing the counts for deaths and expeditions, `left_join()` by season and year, and `replace()` the seasons with no deaths' NA values with zero.

Finally, I will use `ggplot()` and `geom_jitter()` to visualize the data, including a regression line of best fit and coloring points by year. These charts were chosen in order to explore the possibility of a causal relationship between the continuous variables expeditions and deaths. I believe their ability to visually represent three dimensions of continuous information (`year`, `expeditions`, `deaths`,) alongside a best fit line will help in identifying additional patterns.

Analysis:

```
#just your basic filter()
members <-
  members_og |>
    filter(peak_name == "Everest") |>
    filter(year >= 1960)

#and select()
members <-
```

```
  members |>
    select(expedition_id, year, season)

#now we need to collapse observations from individuals into expeditions
#the data now only contains information recorded at the expedition level, so we can treat the additiona
members <-
  members |>
    distinct(expedition_id, .keep_all = TRUE)
```

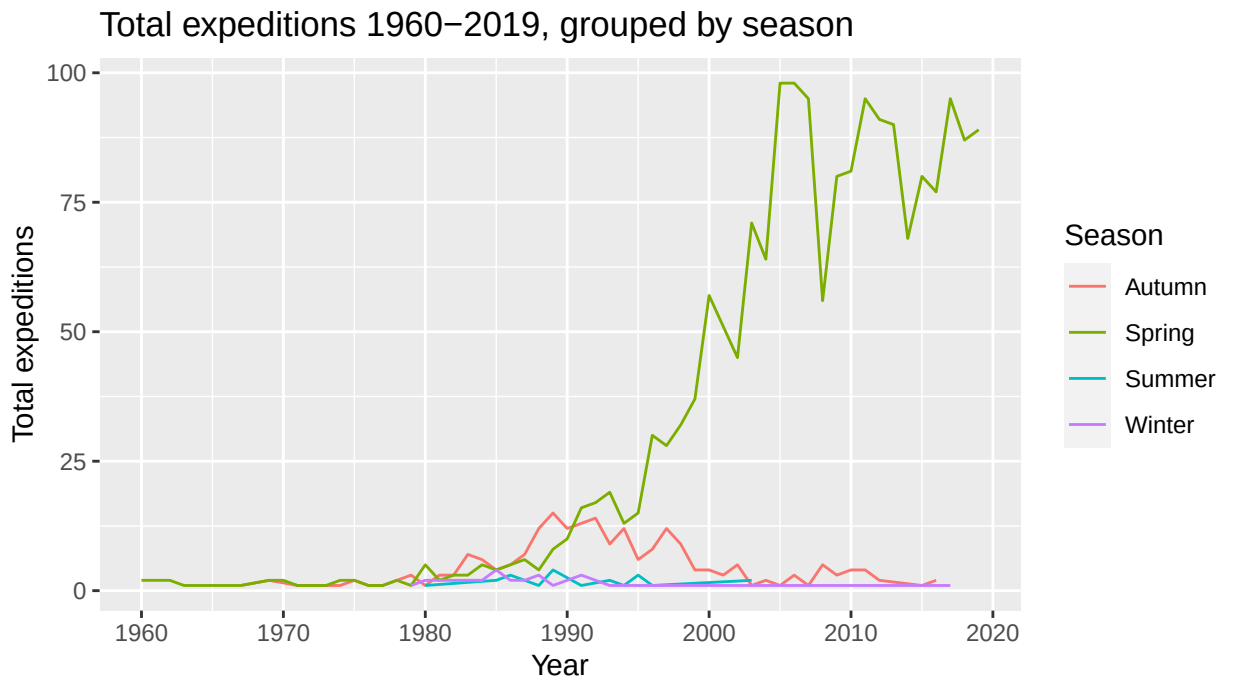
```
#generating a table of expeditions by year and season
table <-
  table(members$year, members$season)
table
```

```
##
##        Autumn Spring Summer Winter
##   1960      0      2      0      0
##   1962      0      2      0      0
##   1963      0      1      0      0
##   1964      0      1      0      0
##   1965      0      1      0      0
##   1966      0      1      0      0
##   1967      1      1      0      0
##   1969      2      2      0      0
##   1970      0      2      0      0
##   1971      1      1      0      0
##   1972      1      1      0      0
##   1973      1      1      0      0
##   1974      1      2      0      0
##   1975      2      2      0      0
##   1976      1      1      0      0
##   1977      1      1      0      0
##   1978      2      2      0      0
##   1979      3      1      0      1
##   1980      1      5      1      2
##   1981      3      2      0      0
##   1982      3      3      0      2
##   1983      7      3      0      2
##   1984      6      5      0      2
##   1985      4      4      2      4
##   1986      5      5      3      2
##   1987      7      6      2      2
##   1988     12      4      1      3
##   1989     15      8      4      1
##   1990     12     10      0      0
##   1991     13     16      1      3
##   1992     14     17      0      2
##   1993      9     19      2      1
##   1994     12     13      1      0
##   1995      6     15      3      0
##   1996      8     30      1      0
##   1997     12     28      0      0
##   1998      9     32      0      0
```

```
##   1999      4     37      0      1
##   2000      4     57      0      0
##   2001      3     51      0      0
##   2002      5     45      0      0
##   2003      1     71      2      0
##   2004      2     64      0      0
##   2005      1     98      0      0
##   2006      3     98      0      0
##   2007      1     95      0      0
##   2008      5     56      0      0
##   2009      3     80      0      0
##   2010      4     81      0      0
##   2011      4     95      0      0
##   2012      2     91      0      0
##   2013      0     90      0      0
##   2014      0     68      0      0
##   2015      1     80      0      0
##   2016      2     77      0      1
##   2017      0     95      0      1
##   2018      0     87      0      0
##   2019      0     89      0      0
```

```
#collapsing observations from individual expeditions into a count by year and season
members <-
  members |>
    count(year, season)
```

```
#generating the graph of expeditions over time by season
members |>
  ggplot(
    aes(
      year,
      n,
      group = season,
      color = season
      )
    ) +
      geom_line() +
      coord_fixed() +
      labs(title = "Total expeditions 1960-2019, grouped by season",
           color = "Season",
           x = "Year",
           y = "Total expeditions"
           ) +
      scale_x_continuous(
          breaks = c(1960, 1970, 1980, 1990, 2000, 2010, 2020)
          ) +
      theme(aspect.ratio=.6)
```

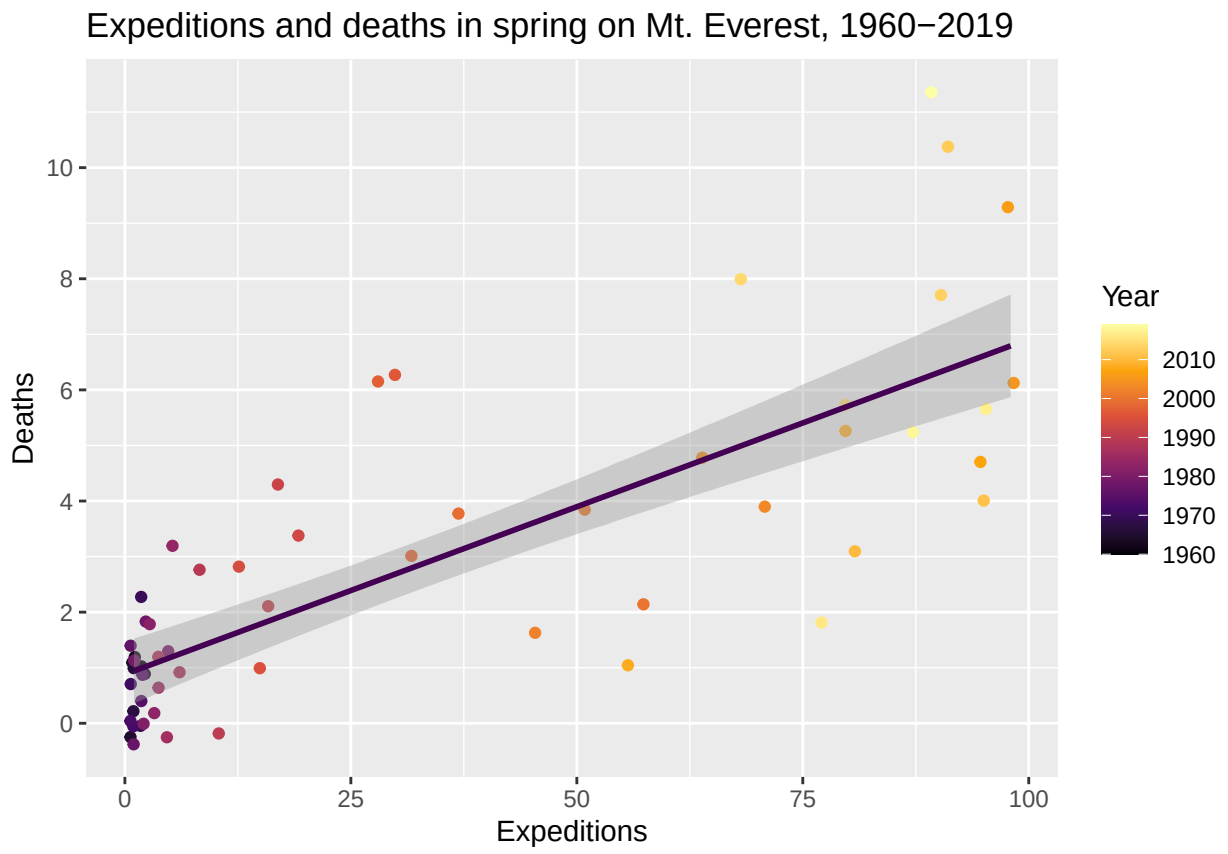



```
#members already contains data about the number of expeditions in each year and season
#so we can rerun the code with only entries containing deaths

#filter()
deaths <-
  members_og |>
    filter(peak_name == "Everest") |>
    filter(year >= 1960) |>
    filter(died == TRUE)

#and select()
deaths <-
  deaths |>
    select(expedition_id, year, season)

#now we need to collapse observations from individuals into expeditions
#the data now only contains information recorded at the expedition level, so we can treat the additiona
deaths <-
  deaths |>
    distinct(expedition_id, .keep_all = TRUE)

#collapsing observations from individual expeditions into a count by year and season
deaths <-
  deaths |>
    count(year, season)
```

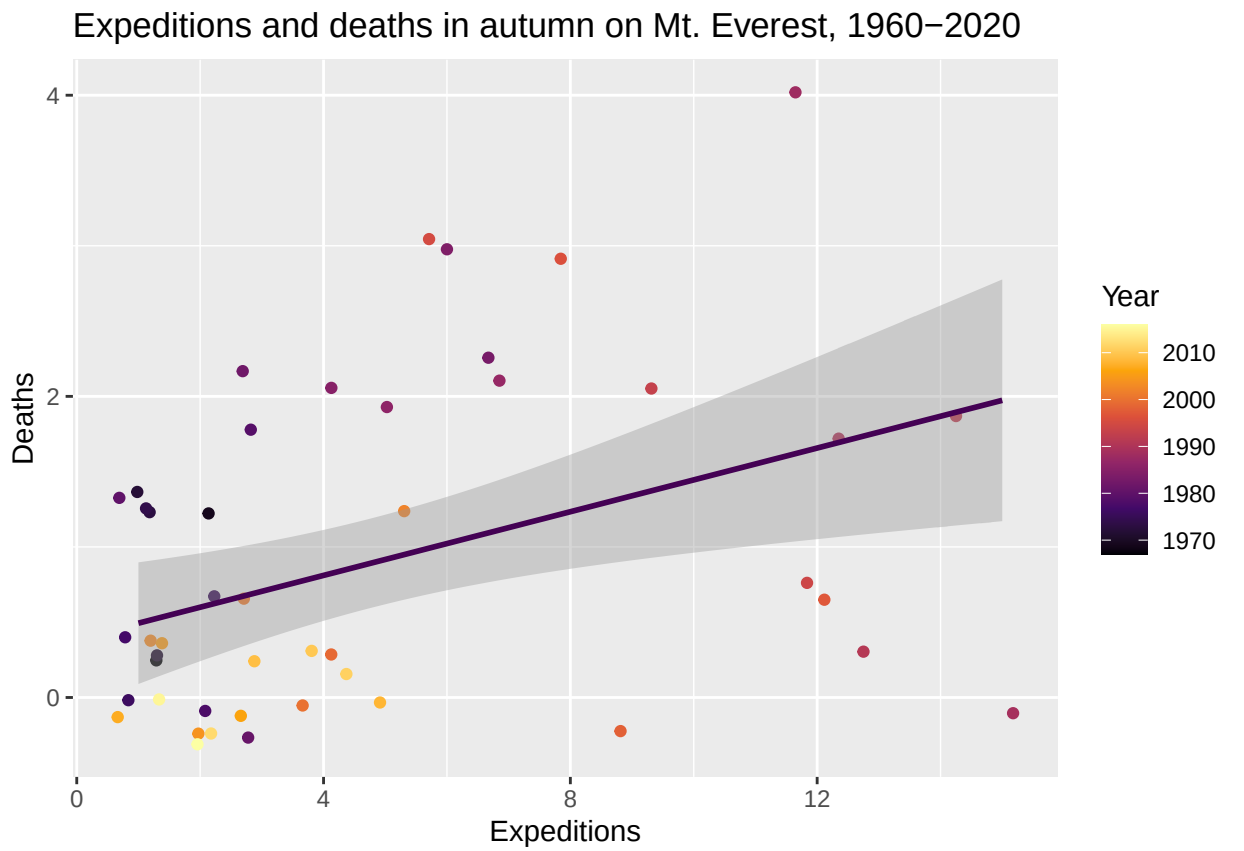
```r
#now we have some renaming to do
deaths <-
  deaths |>
    rename(deaths = n)

members <-
  members |>
    rename(expeditions = n)

#now we merge by year and season
members <-
  left_join(members, deaths, by = c("year", "season"))

#and replace NAs with 0
members <-
  members |>
    replace_na(list(year = NA, season = NA, expedition = NA, deaths = 0))
```

```r
#finally we can generate our graphs, since facet_wrap is a little hard to read
members |>
  filter(season == "Spring") |>
  ggplot(
    aes(
      expeditions,
      deaths,
      color = year
      )
    ) +
      geom_jitter() +
      geom_smooth(method='lm', color = "#440154FF") +
      labs(title = "Expeditions and deaths in spring on Mt. Everest, 1960-2019",
           color = "Year",
           x = "Expeditions",
           y = "Deaths"
           ) +
      scale_y_continuous(
          breaks = c(0, 2, 4, 6, 8, 10, 12),
          labels = c(0, 2, 4, 6, 8, 10, 12)
          ) +
      scale_color_viridis_c(option = "B")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
members |>
  filter(season == "Autumn") |>
  ggplot(
    aes(
      expeditions,
      deaths,
      color = year,
      )
    ) +
      geom_jitter() +
      geom_smooth(method='lm', formula = y~(x^2), color = "#440154FF") +
      labs(title = "Expeditions and deaths in autumn on Mt. Everest, 1960-2020",
           color = "Year",
           x = "Expeditions",
           y = "Deaths"
           ) +
      scale_y_continuous(
          breaks = c(0, 2, 4, 6, 8, 10, 12),
          labels = c(0, 2, 4, 6, 8, 10, 12)
          ) +
      scale_color_viridis_c(option = "B")
```



Discussion:

Once I generated the frequency table of expeditions across seasons and years, it became clear why the death statistics for winter and summer were so prejudiced in favor of falls. The mountain appears to be unapproachable during those periods, with very few expedition attempts being made.

Fall and spring expeditions are much more common, but autumn expeditions hit a local peak for about a decade in the 90s before tapering off as spring expeditions skyrocket in popularity. The local peak, and the incredible boom in summit attempts since 1990, are both readily apparent in the line chart. Clearly spring expeditions have become a sincere favorite. But with so much demand for summit attempts and limited scheduling time in the spring, why have autumn attempts become so rare?

The answer became more clear in the third and fourth visualizations, the scatter plots. Recall from Part 1 that autumn deaths were overwhelmingly caused by avalanches and falls. For the years in the 1990-2000 period, there appeared to be a spike in mortality relative to spring expeditions. This could have been caused by fresh snowpack settling onto the mountain in avalanches, or creating snowdrifts and causing climbers to fall. To determine this to the point of being a statistical certainty I'd need to write more code, and will likely do so later.

Interestingly, the relationship between deaths and expeditions doesn't appear to be linear. I have read elsewhere that Sherpa guides have attempted to make the mountain safer by installing more secure walkways, rope guards, and other permanent safety fixtures. This appears to have created a pattern best suited to a polynomial regression term. This could indicate that Sherpa guides are experiencing economies of scale as tourism interest in Everest grows and investment in the physical infrastructure on Everest makes the ascent safer.