Project 2

Name: Justin Campbell; UT eID: jsc4348

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: *In Part 1 of this project, a data set summarizing Himalayan expeditions pulled from the Himalayan Database will be manipulated through data wrangling and analyzed through data visualizations. The data set is a compilation of records for all treks in the Himalayan mountain range of Nepal from 1905 to 2019, and contains information about properties such as the expedition year, season, expedition member year, and mountain peak to name a few. For the analysis, a breakdown of expedition deaths in each season according to the seven most common causes (totals summed across each of the seasons) will be investigated. This will be accomplished by generating a summary table through the use of ggplot2 functions and the "tidyverse" library applied to the "members" data frame read into the workspace at the beginning of the file. Using the resultant data frame, four pie charts will be generated to depict the relative proportions of each of the deaths for the seven leading death causes across each of the seasons. Four column variables from the data frame will be utilized to answer the question, namely, "year", "season", "peak_name", and "death_cause".*

Approach: *Expanding on the procedure briefly outlined in the "Introduction" section above, the data frame, "members" will be manipulated through data wrangling functions included in the "Tidyverse" library. First, the "filter()" function will be used to filter data from the original data frame so that the resultant data frame contains only data from the year and 1960 and beyond, for the peak "Everest", and for expeditions where the members died. Next, the "select()" function will be used to display two columns in the data frame, namely, the death cause and season for each death. Then, the "count()" function will be used to generate a new column in the data frame with the count of deaths for each pair of death cause and season values. From here, the "arrange()" function will be used to sort the death cause and season value pair counts in descending order. Afterwards, the "pivot_wider()" function will be used to generate a table in wide form that groups the deaths by cause, season, and reports the counts for each category with "N/A" values. From here the "across()" and "replace_na()" functions will be invoked as parameters in the "mutate()" function to replace all N/A values with 0. Lastly, the "group_by()" and "summarize()" functions will be used to display the deaths for the seven* *most common death causes plus an "other" category. It should be noted that the death causes will not be arranged in descending order of frequency according to the total death cause counts, but rather, in ascending alphabetical order for simplicity, and hence, formatting*

Shifting gears to the approach for generating the data visualizations, a vector of the death cause strings will first be extracted from the updated data frame and assigned to a new tibble. Then, a vector of death cause counts for the Spring season will be extracted and assigned to a separate tibble. Thirdly, the two tibbles will be merged to create a new two column tibble with the aforementioned data. Then, the "ggplot" "aes", "geom_col", "coord_polar", "theme_void" and "ggtitle" functions will be used to generate a piechart of relative death cause proportions for the season of Spring. This procedure will then be repeated for the other three seasons.

Analysis:

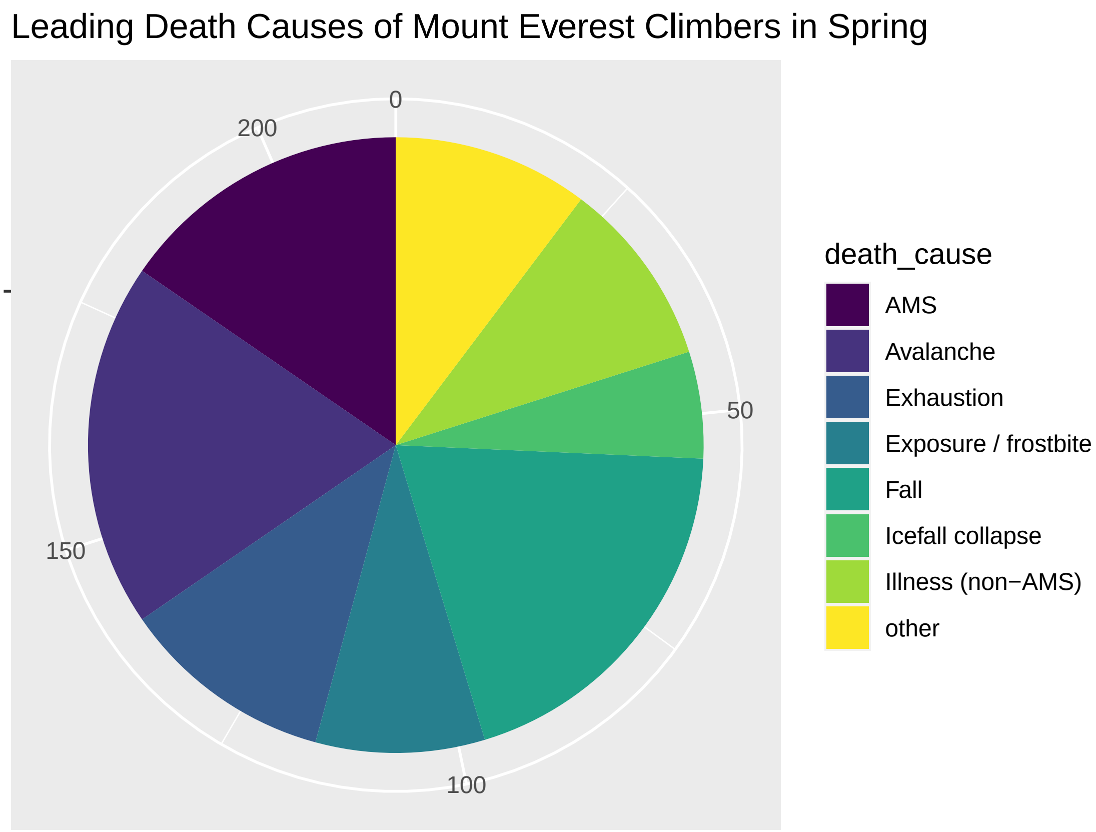
```
# Data Wrangling Section:

# The code in this section uses built-in ggplot2 functions to manipulate the
#"members" dataframe such that it contains 4 columns for each of the respective
#seasons with the count (frequency) of deaths corresponding to the death cause
#in the associated row. The death causes will be sorted in descending order of
#frequency overall (total across all seasons), and the seven most frequent causes
#plus an "other" category will be presented.

members_new_1 <- members %>%

  # Filter data to include "Everest" peak

  filter(year >= 1960, peak_name == "Everest", died == TRUE) %>%

    # Report each individual death by the season and death cause

    select(death_cause, season) %>%

      # Count and display death frequencies according to season and death cause

      count(death_cause,season)  %>%

        # Arrange death cause and season pair counts in descending order

        arrange(desc(n))  %>%

          # Group deaths by cause and season and report counts for each category

          pivot_wider(names_from = "season", values_from = "n")  %>%

            # Replaces NA values with 0

             mutate(across(everything(), ~replace_na(.x, 0))) %>%
              group_by(death_cause = ifelse(row_number() < 8, death_cause, "other")) %>%
                summarize(across(everything(), sum))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

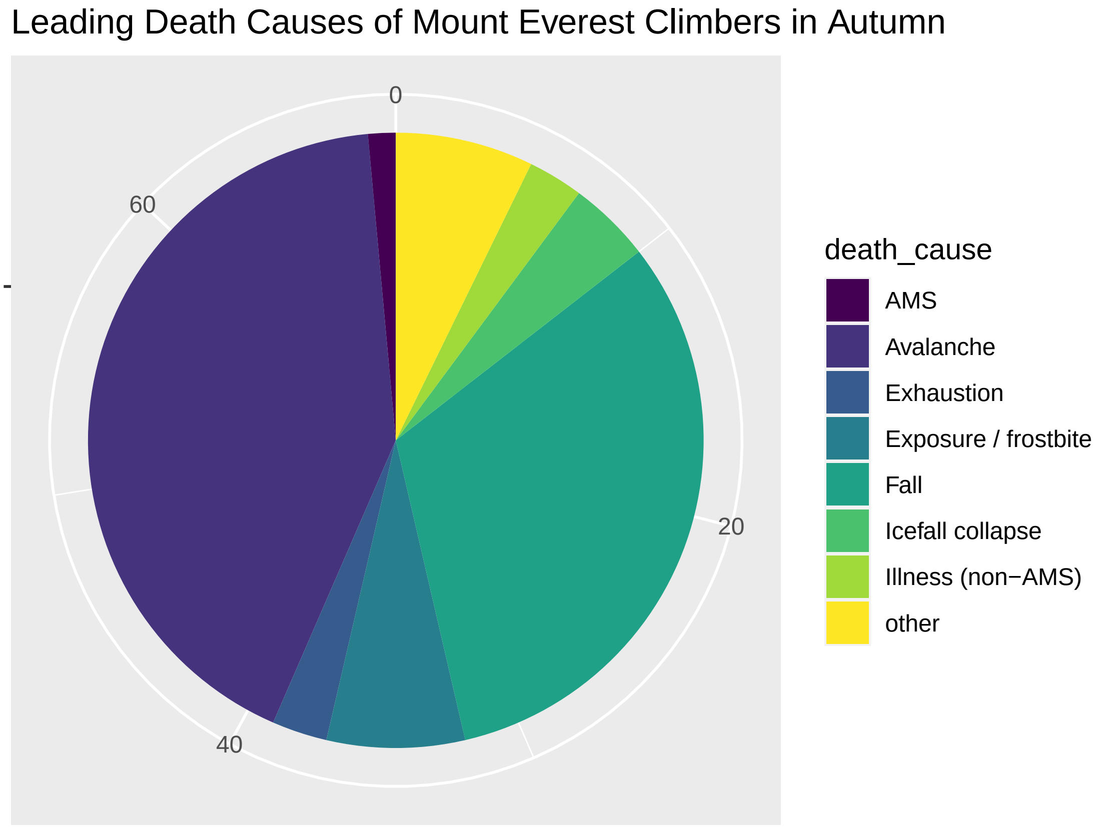
```
# Output the new data frame with death counts grouped according to death cause
# and season for the 7 most common death causes plus the "other" category

members_new_1
```

```
## # A tibble: 8 x 5
##   death_cause          Spring Autumn Winter Summer
##   <chr>                 <dbl>  <dbl>  <dbl>  <dbl>
## 1 AMS                      33      1      1      0
## 2 Avalanche                41     29      0      0
## 3 Exhaustion               24      2      0      0
## 4 Exposure / frostbite     19      5      0      0
## 5 Fall                     42     22      5      1
## 6 Icefall collapse         12      3      0      0
## 7 Illness (non-AMS)        21      2      0      0
## 8 other                    22      5      1      0
```

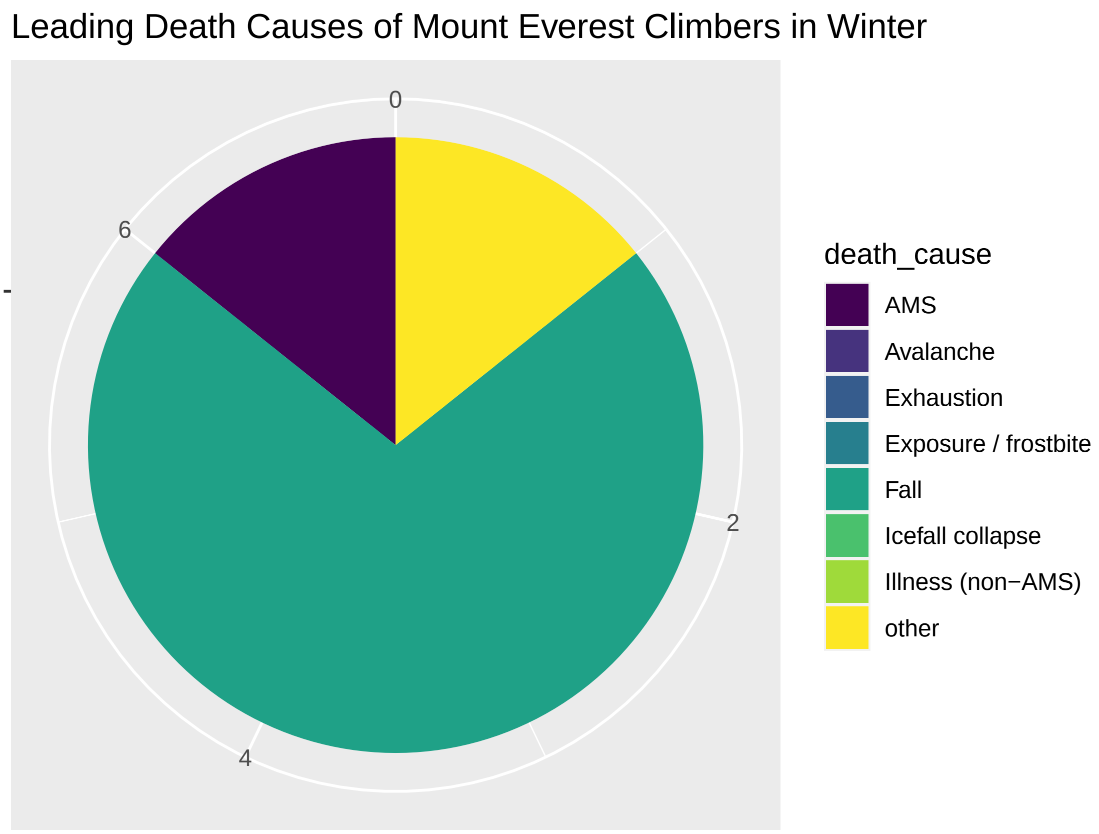
```
# Plotting Section:

# The code in this section uses ggplot2 functions provided in the "Visualizing
#Proportions" lecture to generate four pie charts depicting the relative
#proportions of death causes for each associated season.

# Extract death cause vector from "members_new" dataframe

death_cause <- members_new_1 %>% select(death_cause)

# Extract vector of death cause counts for "Spring" season

Spring <- members_new_1 %>% select(Spring)

# Generate new tibble with "death_cause" and "Spring" vectors

spring_death_cause <- tibble(death_cause, Spring)

# Use "ggplot", "aes", "geom_col", "coord_polar", "theme_void" and "ggtitle"
#functions to generate piechart of relative death cause proportions for spring

ggplot(spring_death_cause) + aes(Spring, "", fill = death_cause) + geom_col() +
  coord_polar() + theme(axis.title = element_blank()) + scale_fill_viridis_d() +
    ggtitle("Leading Death Causes of Mount Everest Climbers in Spring")
```



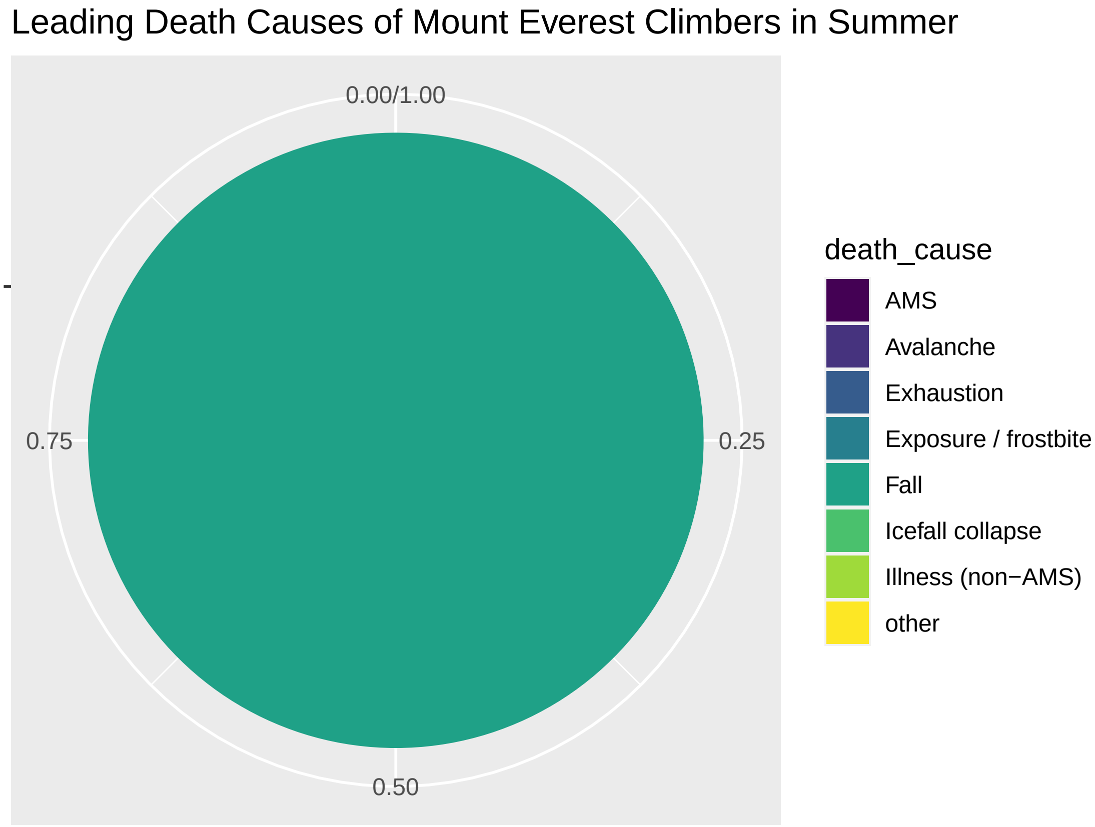
```r
 # Extract vector of death cause counts for "Autumn" season

Autumn <- members_new_1 %>% select(Autumn)

# Generate new tibble with "death_cause" and "Autumn" vectors

autumn_death_cause <- tibble(death_cause, Autumn)

# Use "ggplot", "aes", "geom_col", "coord_polar", "theme_void" and "ggtitle"
#functions to generate piechart of relative death cause proportions for autumn

ggplot(autumn_death_cause) + aes(Autumn, "", fill = death_cause) + geom_col() +
  coord_polar() +  theme(axis.title = element_blank()) + scale_fill_viridis_d()+
    ggtitle("Leading Death Causes of Mount Everest Climbers in Autumn")
```



```
# Extract vector of death cause counts for "Winter" season

Winter <- members_new_1 %>% select(Winter)

# Generate new tibble with "death_cause" and "Winter" vectors

winter_death_cause <- tibble(death_cause, Winter)

# Use "ggplot", "aes", "geom_col", "coord_polar", "theme_void" and "ggtitle"
#functions to generate piechart of relative death cause proportions for winter

ggplot(winter_death_cause) + aes(Winter, "", fill = death_cause) + geom_col() +
  coord_polar() + theme(axis.title = element_blank()) + scale_fill_viridis_d() +
    ggtitle("Leading Death Causes of Mount Everest Climbers in Winter")
```



```r
 # Extract vector of death cause counts for "Summer" season

Summer <- members_new_1 %>% select(Summer)

# Generate new tibble with "death_cause" and "Summer" vectors

summer_death_cause <- tibble(death_cause, Summer)

# Use "ggplot", "aes", "geom_col", "coord_polar", "theme_void" and "ggtitle"
#functions to generate piechart of relative death cause proportions for summer

ggplot(summer_death_cause) + aes(Summer, "", fill = death_cause) + geom_col() +
  coord_polar() + theme(axis.title = element_blank()) + scale_fill_viridis_d() +
    ggtitle("Leading Death Causes of Mount Everest Climbers in Summer")
```



Leading Death Causes of Mount Everest Climbers in Summer

Discussion: *In referencing the first of the four pie charts output in the analysis section, we see that the three leading causes of death in the season of Spring over the time frame of 1960 to 2019 are "Fall", "Avalanche", and "AMS (Acute Mountain Sickness)" respectively, while the least frequent causes are "Icefall collapse", "Exposure / frostbite", and "Illness (non-AMS)" respectively. As it pertains to the leading causes of death, it is reasonable that the two most frequent death causes in the season of Spring are falling, and avalanches. This can largely be attributed to an increase in temperature during the Spring season months that ultimately melts much of the snow and ice. This in turn, creates areas of loosely packed snow and liquid runoff that reduces traction and can cause mountain hikers to loose their footing and fall, or cause avalanches from accumulation of the runoff. Acute Mountain Sickness is another leading cause of death during the warming months of Spring because hikers tend to trek quickly during this transitional period in temperature because visibility at this time of year improves from the Winter. When, hikers ascend through altitude levels too quickly, their body does not have a sufficient amount of time to adjust to the reduced oxygen levels which causes AMS. It is reasonable that icefall collapse, exposure/frosbite, and non-AMS related illness are the least frequent causes during this time of year as the temperatures begin to increase to more comfortable levels.*

Looking at the second figure, we see a considerably different array of proportions for each of the seven leading causes of death. In particular, the three leading causes are "Avalanche", "Fall", and "Exposure / Frostbite", while the three least frequent causes are "Exhaustion", "Icefall collapse", and "Illness (non-AMS)". It is plausible that avalanches are the leading cause of death in Autumn given that, (like in the Spring) temperatures are rapidly changing relative to other times of the year, which changes the packing of the ice and snow causing loose precipitation to accumulate into avalanches. What comes with avalanches, are falls associated with reduced traction and footing of mountain climbers. It is interesting to note that these two leading causes account for nearly 75% of the deaths. It is reasonable that exhaustion and icefall collapse, and non-AMS related illness are the least frequent causes as the temperatures are cooling down, and thus, not extreme during these months of the year.

In referencing the third figure, we see that only three causes of death are reported, namely, "Fall", "AMS", *and "Other". Additionally, we note that these counts are very small with respect to the counts of the leading death causes studied in the seasons of Spring and Autumn. This is likely because the weather conditions in the winter are largely uniform in temperature and precipitation, and thus, there are not many changes to the features of the peak that would affect climbing safety (such as loose snow/ice, avalanches, etc.). It is then reasonable that falling is the leading cause because regardless of the time of year, there will always be steep terrain that inexperienced climbers traverse and either don't have the strength, balance, or footing to avoid falling.*

Lastly, in referencing the pie chart for the season of Summer, we see that the only cause of death on Mount Everest from 1960 to 2019 was from a fall that a single mountain climber experienced. It is reasonsble that the number and causes of death in the season of Summer would be the least among the four seasons as the temperatures are the warmest and relatively uniform with time. This means that temperatures change very little over the Summer months which reduces the likelihood of avalanches, melting ice/snow, and other changing properties.

Part 2

Question: *How does the distribution of maximum height climbed for those who died while climbing Mount Everest vary according to the seasons of Autumn and Spring, and according to sex?*

Introduction: *In Part 2 of this project, the "members" data frame analyzed in Part 1 will again be manipulated through data wrangling and important properties analyzed through data visualizations. In particular,. for the analysis, the distribution of the maximum height climbed for climbers who died while on Mount Everest will be investigated according to the most dangerous climbing seasons of the year (Spring and Autumn), and according to sex. This will be accomplished by first filtering out the variables "peak_name", "sex", "season" , "died", and eliminating all rows whose "highpoint_metres" value is "N/A". The tibble will then be simplified to include only the "sex", "season", and "highpoint_metres" variables. These two steps will be performed through the use of ggplot2 functions and the "tidyverse" library applied. Using the resultant data frame, four histograms will be generated to depict the distributions of high point climbed in metres according to season and sex.*

Approach: *The first step that will be performed in the data wrangling section of the analysis involves filtering the data frame to include only climbs on Mount Everest, in the seasons of Spring or Autumn, climbs that resulted in a death, and climbs whose high point in metres are provided. This will be accomplished using the "filter()" function. Then, the resulting data frame will be operated on using the "select()" function to eliminate all remaining entries from the data frame with exception to those that correspond to the "highpoint_metres", "season", and "sex" variables. Lastly, the resultant data frame will be output to view the simplified properties.*

After generating a simplified data frame in the data wrangling section of the analysis, the "ggplot()", and "geom_histogram()" functions will be used to generate a subplot with four histograms of the high point climbed according to the seasons of Spring and Autumn, and according to sex. This will be accomplished first by coding the "ggplot()" function with the updated data frame as the first argument, and the "aes()" function call with "highpoint_metres" and "after_stat(count)" as its two parametes for the second argument. The "+" operator will be used to continue the plot generation sequence of commands allowing for the use of "geom_histogram()" where a bin width of approximately 500, and a fill color scheme of "skyblue" will be used. Again adding the "+" operator, the "facet_wrap()" function will be invoked to group the distributions according to season and sex. Next, the "theme()" function will be used to define a visually appropriate theme to the canvas of the figure window. Next, the"ggtitle()" function will be used to add a descriptive title to the figure window. Lastly, the "labs()" function will be used to add meaningful labels to the figure axes.

Analysis:

```
# Data Wrangling Section:

# The code in this section uses built-in ggplot2 functions to manipulate the
#"members" data frame such that it contains the high point climbed in metres for
```

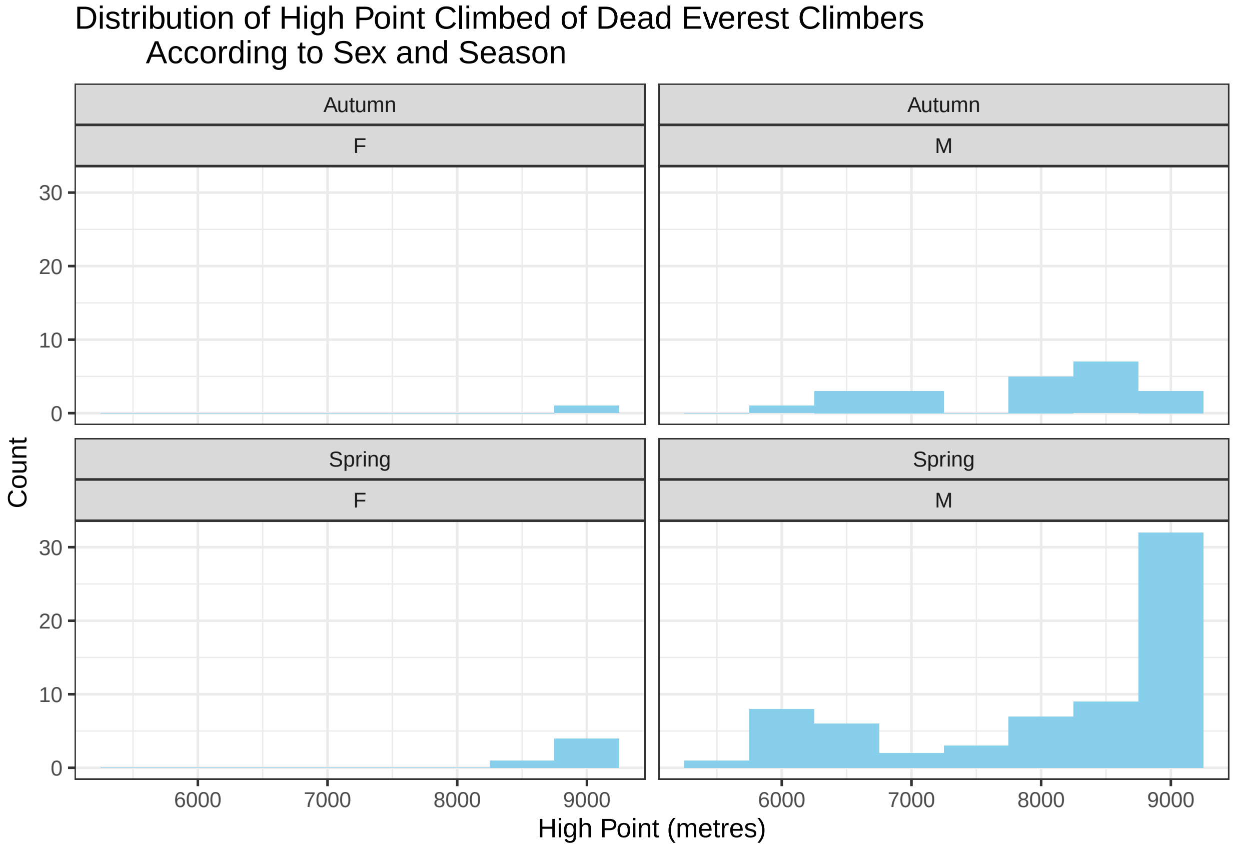
```r
#climbers that died while trekking Mount Everest in either the season of Spring
#or Autumn according to their sex

members_new_2 <- members %>%

   # Filter data to include "Everest" peak and deaths, and spring and autumn
   # records only

  filter(peak_name == "Everest", season == c("Spring", "Autumn"), died == TRUE,
  !is.na(highpoint_metres)) %>%

    # Report each individual death by the season, sex of climber, and high point
    #climbed in meters

    select(highpoint_metres, season, sex)
```

```
## Warning in season == c("Spring", "Autumn"): longer object length is not a
## multiple of shorter object length
```

```r
# Output the new data frame with high point climbed in metres, season of climb,
# and sex of climber

members_new_2
```

```
## # A tibble: 96 x 3
##    highpoint_metres season sex
##               <dbl> <chr>  <chr>
##  1             5800 Spring M
##  2             6400 Autumn M
##  3             6650 Spring M
##  4             8850 Autumn F
##  5             8600 Autumn M
##  6             8850 Autumn M
##  7             8850 Spring M
##  8             8400 Autumn M
##  9             8140 Autumn M
## 10             7600 Spring M
## # ... with 86 more rows
```

```r
# Plotting Section: In this code block, the "ggplot()" and "geom_histogram()"
#functions will be used to generate a subplot with four histograms that depict
#the distribution of high point climbed in metres of Mount Everest climbers who
#died according to the seasons of Spring and Autumn and according to sex.

# Initialize figure window with "highpoint_metres" as quantitative variable and
#"count" as a measure of the frequency

ggplot(members_new_2, aes(highpoint_metres,  y = after_stat(count))) +

  # Initialize histogram calls with binwidth of 500 and fill color of "skyblue"

   geom_histogram(binwidth = 500, fill = "skyblue") +

    # Generate subplot of four histograms according to sex and season
    facet_wrap(vars(season, sex))  + theme_bw(10) +

        # Define descriptive title

        ggtitle("Distribution of High Point Climbed of Dead Everest Climbers
        According to Sex and Season") +

           # Define meaningful axis labels

           labs(x = "High Point (metres)", y = "Count")
```



Discussion: *In referencing the first subplot on the figure window, namely, the distribution of high point climbed amongst female Everest climbers who died while trekking in the season of Autumn, we see just a single bin of data with a frequency of two. This indicates that very few female mountain climbers have died while climbing Mount Everest in the season of Autumn,and that those that have, ascended to a relatively high altitude of between 8750 and 9250 meters. Shifting gears to the subplot at the top right, there is no distinct trend as it pertains to the shape of the data, but that the center of the data appears to be around 8000 meters, and that the highest bin frequencies are located at the upper-end of the distribution, namely, bins corresponding to 7750-8250, and 8250-8750. This suggests that the highest number of deaths of male climbers in the season of Autumn occured when the maximum ascended height was between 7750 and 8750 meters, namely at the upper-end of the spectrum as was the case for the females during the same seasonal period. It is also interesting to note that there is a gap between 7250 and 7750 suggesting that the conditions at this altitude during Autumn is safest in the spectrum, while the conditions at the altitude of between 8250 and 8750 meters in the least safe. An important distinction that should be made is that the number of males that died on Mount Everest in the season of Autumn is much greater than that of the number of females.*

Looking at the bottom left subplot, we see a distribution that is very similar to that of the top-left subplot, namely, of the females that died in ascending the mountain during the Spring season, nearly all ascended to a maximum height of between 8750 and 9250 meters. However, the frequency of the deaths is noticeably larger than in the Autumn season. Lastly, looking at the figure in the bottom-right, we see that the distribution is skewed at its lower end. Additionally, the highest bin frequency by far is at the upper-end of the distribution corresponding to a high point in meters of between 8750 and 9250. Similar to the features shown for the males in the season of Autumn, the conditions at the altitude of between 8750 and 9250 for both males and females in the season of Spring are the most dangerous. The conditions at altitudes between 5250 and 5750 and between 6750 and 7250 for the males during this season appear to the safest, in contrast to the range of 8250 to 8750 for the females. Although, the latter cannot be concluded with great certainty considering the small number of death records for females in this seasonal time period.

Overall, the subplots show that climbing Mount Everest in the Autumn months is historically moderately safter than climbing in the Spring months for both females and males. This is largely due to the increase in temperatures in the Spring months that melt the ice and snow, reducing climbers footing and traction. It is also shown that the vast majority of climbers that have died while hiking Mount Everest are male, and it is plausible to infer that this trend results from a disproportionately large number of male climbers that have historically climbed Mount Everest (although this cannot be directly shown from the data).