Project 2

Jason Flores UT EID: *jf36995*

This is the dataset you will be working with:

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

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:

We are working with the `members` dataset which contains records of all the individuals who partook in the Himalayan expeditions from 1905 through Spring 2019 covering more than 465 peaks in Nepal. In this data set `members`, each row contains the individual expedition number and route they took, showing their expedition, member and peak id. It has 17 additional variables that explains the year and season of the expedition, followed by the sex of the individual and age and what citizenship they have. Furthermore, it has information of the role they did during the expedition and if they were hired. It also has information of the elevation of the highpoint in meters and if the expedition was a success. Lastly, it also explains if oxygen was used and if the expedition was done solo and whether the expedition had any casualties describing the cause of death, height of death, or type of injury and height of injury.

To answer question 1, we use data set `members` and the variables we will be working on is the (column `death_cause`) to determine the seven most death causes. Furthermore, the data will contain the season the death occurred in i.e (column `season`) and a pie chart will be plotted to view the ratio of the deaths. Overall, a table containing the respectable variables will be formed where the pie chart of that table is viewed.

Approach:

My approach in answering question 1 is to create a table using logic indexing (calling it `sumTable`) containing six columns that has `death_cause`, Spring, Summer, Autumn, Winter, and Total. First, we have to filter the data starting in 1960 and peak name to Everest, then we use the`fct_lump_n` to group the remaining death causes as other. Afterwards we use the `count` function to find all the observations of `death_cause` and `season`. Now, we use the `pivot_wider` function to view the raw deaths of each cause by season. Then, we change all Na's value to zero using `mutate(across(everthing))`. Now, we can use a `mutate` function to

find the total using `rowSums`. Then we can arrange the data by total where we remove the row containing 0 as the name using `slice`. Lastly, we use `select` to reorder the dataframe in the appropriate order wanted.

After establishing the summary table, we call the logic indexing where we simply use a ggplot to view it as a pie chart using the function `geom_arc_bar`. However, since it is asking to facet by season we need the dataset in its long form. Therefore, the function `pivot_longer` is used.

Analysis:

```
#creating summary table, calling it sumTable
sumTable <- members %>%
   filter(peak_name == "Everest", year >= 1960) %>%
#Assigning other causes as other
  mutate(
     death_cause = fct_lump_n(fct_infreq(death_cause), 7, other_level = "Other")
     )%>%
#Count
  count(death_cause, season) %>%
#Pivot Wider
  pivot_wider(names_from = season, values_from = n) %>%
#changing NA's to 0
  mutate(
    across(everything(), ~replace_na(.x, 0))
  ) %>%
#Finding total deaths for each cause
  mutate(
    Total = rowSums(across(where(is.numeric)))
  ) %>%
#slicing data frame to not view the NArow
 slice(1:8) %>%
#using select to have dataframe in order wanted
  select(death_cause,Spring,Summer,Autumn,Winter,Total) %>%
  unique()
```

```
## Warning: Problem with `mutate()` input `..1`.
## i invalid factor level, NA generated
## i Input `..1` is `across(everything(), ~replace_na(.x, 0))`.
```

```
## Warning in `[<-.factor`(`*tmp*`, !is_complete(data), value = 0): invalid factor
## level, NA generated
```

```
print(sumTable)
```

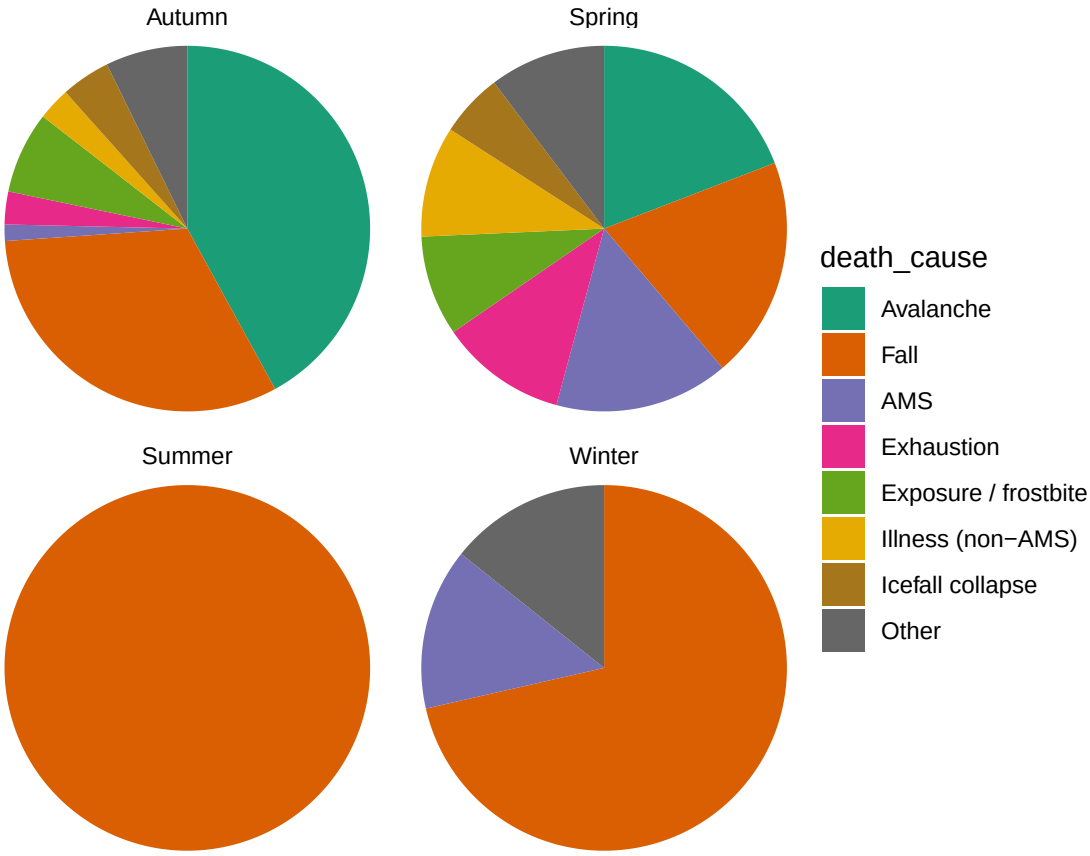
```
## # A tibble: 8 x 6
##   death_cause          Spring Summer Autumn Winter Total
##   <fct>                 <dbl>  <dbl>  <dbl>  <dbl> <dbl>
## 1 Avalanche                41      0     29      0    70
## 2 Fall                     42      1     22      5    70
## 3 AMS                      33      0      1      1    35
## 4 Exhaustion               24      0      2      0    26
## 5 Exposure / frostbite     19      0      5      0    24
## 6 Illness (non-AMS)        21      0      2      0    23
## 7 Icefall collapse         12      0      3      0    15
## 8 Other                    22      0      5      1    28
```

```
#Changing dataframe to longer format
piechart <- sumTable %>%
```

```
   select(death_cause,Spring,Summer,Autumn,Winter) %>%
      pivot_longer(cols = c(Spring:Winter), names_to = "season", values_to = "Count")
  print(piechart)
```

```
## # A tibble: 32 x 3
##    death_cause season Count
##    <fct>       <chr>  <dbl>
##  1 Avalanche   Spring    41
##  2 Avalanche   Summer     0
##  3 Avalanche   Autumn    29
##  4 Avalanche   Winter     0
##  5 Fall        Spring    42
##  6 Fall        Summer     1
##  7 Fall        Autumn    22
##  8 Fall        Winter     5
##  9 AMS         Spring    33
## 10 AMS         Summer     0
## # ... with 22 more rows
```

```
#graph of datafram piechart as a pie chart
   ggplot(piechart) +
  aes(
    x0 = 0, y0 = 0,
    r0 = 0, r = 1,
    amount = Count,
    fill = death_cause
  ) +
  geom_arc_bar(stat = "pie", color = NA) +
     scale_fill_brewer(palette="Dark2") +
     facet_wrap(~(season)) +
  coord_fixed() +
theme_void()
```



Discussion:

It was determined that the seven most caused deaths overall in the expedition to Mt.Everest peak from 1960 to Spring 2019 was by avalanche (70), falling (70), AMS (35), exhaustion (26), exposure/frostbite (24), succumbing to illness (23), icefall collapse (15), and other (28). Furthermore, by viewing the pie charts it was concluded that death by falling is seen across all seasons and the only death cause occurring in the summer. Additionally, all 8 death causes are seen in the spring and autumn expeditions. It can be inferred that more expeditions occur at that time of season. In conclusion, all 8 death causes are seen in the Spring and Autumn season 3 death causes in the winter and only 1 death cause in the summer, and most recorring death is death by avalanche or falling.

Part 2

Question:

Using the three highest peaks in the Himalayan expedition what is the overall deaths since 2000?

Introduction:

As mentioned in part 1, we will be working with the `members` data set. To answer question 2 the variables that we will be focusing on are (column `peak_names`) using *Everest*, *Kangchenjunga*, *Lhotse* and (column `died`) which states if any deaths occured. After obtaining the data frame, a bar plot will be used to visualize the amount of deaths in each peak.

Approach:

The approach is similar to question 1, a summary table called `peakTable` is created where first the `filter` function was used to filter the data by year 2000 and the three peaks investigated being *Everest*, *Kangchenjunga*,

Lhotse. After filtering the data, the `died` column *FALSE* and *TRUE* was changed to *No* and *Yes* respectively using `mutate` and `if_else`. From there the `count` function was used to count the number of deaths across the three peaks. Now, to visualize the table better the `pivot_wider` function is used to separate the three peaks as columns. Then another `mutate` function is used to find the overall deaths where is shown in a column labeled as *Total*. Lastly, another `mutate` function is used to view the percentage of the overall deaths.

After creating the `peakTable`, the data frame is changed to a longer format using `pivot_longer` to create a bar plot using `geom_col` to view the deaths across all three peaks.

Analysis:

```
peakTable <- members %>%
#Filtering by year 2000 and the 3 highest peaks("Everest","Kangchenjunga","Lhotse")
  filter(peak_name %in% c("Everest","Kangchenjunga","Lhotse"), year >= 2000) %>%
#changing died to yes and no
mutate(
    Died = if_else(died == "FALSE","No", "Yes")
  ) %>%
#counting the deaths and peak names
 count(Died,peak_name) %>%
#changing data to wide format
pivot_wider(names_from = peak_name, values_from = n) %>%
#adding total
mutate(
    Total = rowSums(across(where(is.numeric)))
  ) %>%
#finding the percent of deaths from the 3 expeditions
  mutate(
    percentage = 100*Total/sum(Total)
  )
print(peakTable)
```

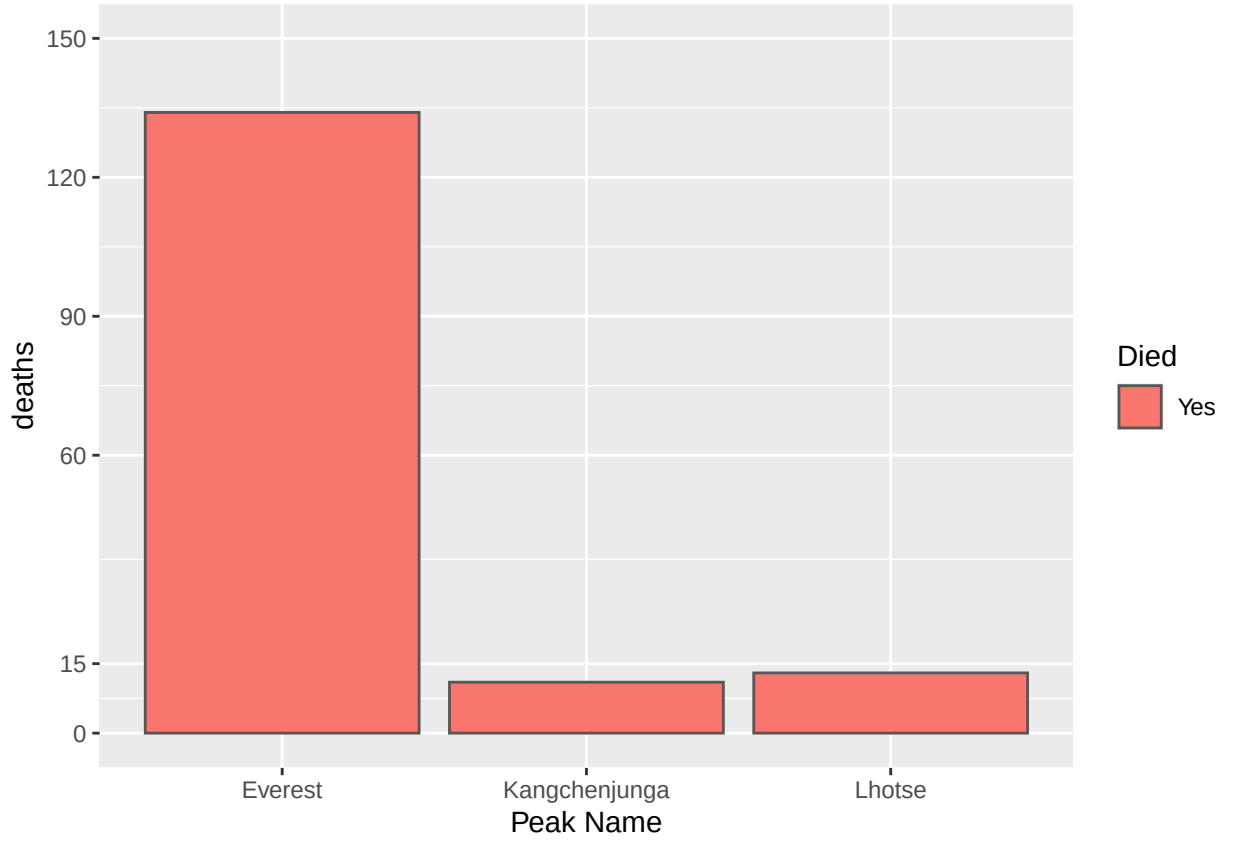
```
## # A tibble: 2 x 6
##   Died  Everest Kangchenjunga Lhotse Total percentage
##   <chr>   <int>         <int>  <int> <dbl>      <dbl>
## 1 No      15015           617   1734 17366     99.1
## 2 Yes       134            11     13   158      0.902
```

```
#changing data peakTable back to long format
peakplot <- peakTable %>%
  select(Died,Everest,Kangchenjunga,Lhotse) %>%
  pivot_longer(cols = c(Everest:Lhotse), names_to = "PeakName", values_to = "Count") %>%
  slice(4:6)
  print(peakplot)
```

```
## # A tibble: 3 x 3
##   Died  PeakName      Count
##   <chr> <chr>         <int>
## 1 Yes   Everest         134
## 2 Yes   Kangchenjunga    11
## 3 Yes   Lhotse           13
```

```
# bar plot of deaths for each peak
  ggplot(
    peakplot,
    aes(PeakName,Count, fill = fct_reorder(Died, -Count))) +
  geom_col(color = "gray34", position = "dodge") +
```

```
labs(fill = "Died")+
xlab("Peak Name")+
scale_y_continuous(
name = "deaths",
limits = c(0, 150),
breaks = c(0, 15, 60,90,120,150),
labels = c("0", "15","60","90","120","150")
)
```



Discussion:

Looking at the expeditions in Mt.Everest, Mt.Kangchenjunga and Mt. Lhotse since 2000 to Spring 2019 the number of deaths that occurred was 158. Furthermore, it appears that the majority of the expeditions occurred in Mt.Everest followed by Mt.Lhotse and Mt.Kangchenjunga. The Mt.Everest expedition had 134 deaths, Mt.Kangchenjunga had 11 deaths, and Mt. Lhotse had 13 deaths. Out of all the expeditions observed in these 3 peaks the overall deaths was 0.902 %