Project 2

Joshua Bryer jeb5533

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 76519 observations from records for all individuals who participated in expeditions to the Nepalese Himalayas from 1905 through Spring 2019 to more than 465 significant peaks in Nepal. In this dataset, each row corresponds to a member on an expedition, and there are twenty-one columns providing information about the peak name, the year of the expedition, the season, the age of the member, citizenship of the member, whether they used oxygen, and whether they were successful.*

To answer the question of Part 1, we will work with two variables, the season(season)and if they died how they died(death _cause). The season is provided as a categorical value. The death causes are listed as categorical variables each describing the manner in which they died, for instance "Avalanche" means they died in an avalanche.

Approach: *Our approach is to show the distributions of causes of death versus the season using a pie chart geom_col()+coord_polar. We separate deaths by cause and seasons, because seasons have much different weather patterns and likely causes of death therefore must be considered separately. Using facet_wrap and pie charts allow us to quickly see the proportion of the seven most common types of member deaths.*

One limitation of the pie is that they don't show us how many observations fall into the different categories. Therefore, we will create a summary table including the number of deaths by cause in each season using the function pivot_wider(). Jointly, these two plots will allow us to answer the question.

Analysis:

```
# This code filters the dataset for the ranges we are looking at, selects our variables, and then gives
df<-members%>%
  filter(year>=1960)%>%
```

```
  filter(peak_name=="Everest")%>%
  dplyr::select(season,death_cause)%>%
  na.omit()%>%
  mutate(death_cause = fct_lump_n(fct_infreq(death_cause), 7))%>%
  count(death_cause,season)
df
```

```
## # A tibble: 20 x 3
##    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
```

```
# This code uses recoded values of the "members" dataset to create a table showing the total number of
table<-df%>%
   pivot_wider(names_from = "season", values_from = "n")%>%
   mutate(across(everything(), ~replace_na(.x, 0)))
```

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

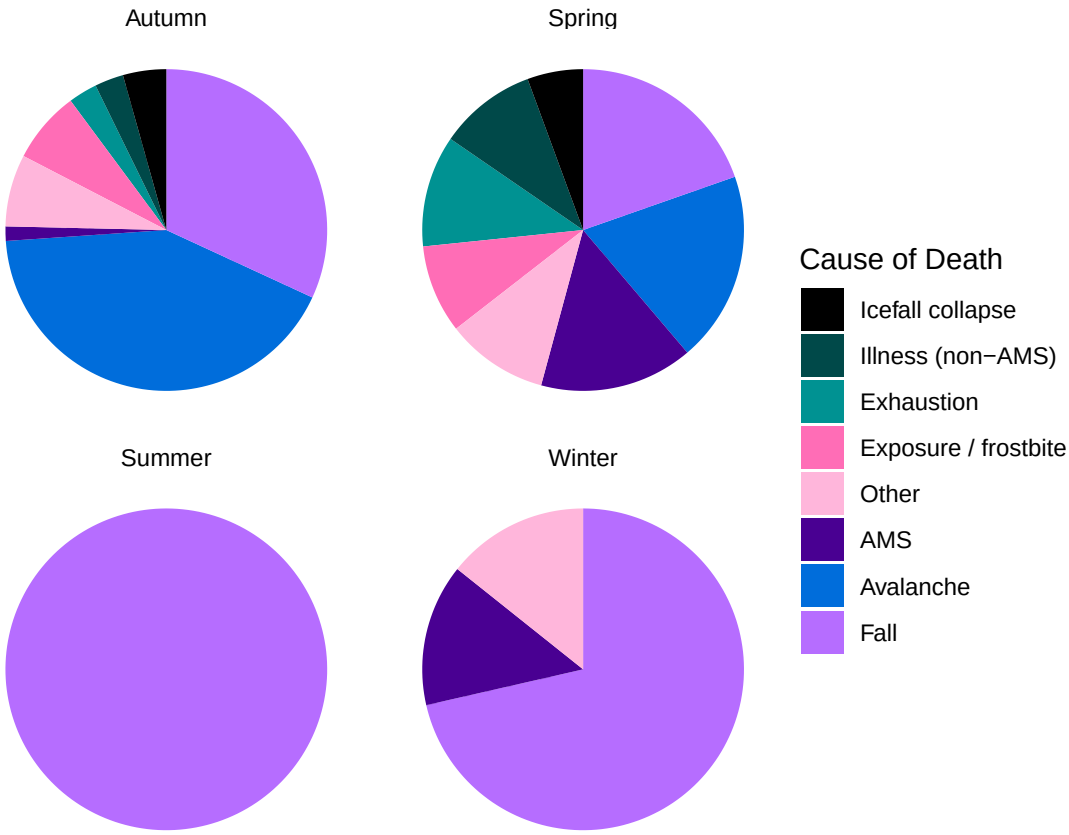
```
## Warning in `[<-.factor`(`*tmp*`, !is_complete(data), value = 0): invalid factor
## level, NA generated
```

```
table
```

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

```r
# This code recodes the column died and then lumps the column death cause and indexes it all into the d
library(paletteer)
pidata<-df%>%
  group_by(season)%>%
  mutate(prop=n/sum(n))

pidata%>%
  ggplot(aes(prop,"YY",fill=fct_reorder(death_cause,prop)))+
  geom_col()+
  theme_void()+
  coord_polar()+
  facet_wrap(vars(season))+
  labs(fill="Cause of Death")+
  paletteer::scale_fill_paletteer_d("colorBlindness::paletteMartin")
```



Discussion: *For distribution of cause of death,season appears to have a major effect on the likely cause of death. We can see this by comparing the distribution of death causes between the seasons, we see that they are majorly shifted relative to each other and list different distributions of each cause per season. We can clearly see the pattern but we would have to run a statistical analyses to determine whether the observed patterns are statistically significant.*

Part 2

Question: *Looking at expeditions how does the frequency of success on the top 3 most frequented Mountains break down versus all the rest?*

Introduction:

To answer the question of Part 1, we will work with two variables, success(success) and which peak they were ascending(peak_name). The success is provided as TRUE or FALSE value. The death peak names are listed as a categorical variable, each indicating a seperate peak, for instance "Everest" means they were climbing Mount Everest.

Approach: *Our approach is to show the distributions of successes versus the top three most attempted mountains using a bar chart geom_col(). We separate successes by the mountain, because each mountain provides very different challenges and must be considered separately. Using facet_wrap and bar charts allow us to quickly see the proportion of successes on the top three most attempted mountain ascents.*

One limitation of the bar chart is that it does not show us how many observations fall into the different categories. Therefore, we will create a summary table including the number of successes per each mountain using the function pivot_wider(). Jointly, these two plots will allow us to fully answer the question.

Analysis:

```
# This code selects our variables, removes unwanted levels in peak_name, and counts the occurences
p2<-members%>%
  dplyr::select(success,peak_name)%>%
  na.omit()%>%
  mutate(peak_name = fct_lump_n(fct_infreq(peak_name), 3))%>%
  count(success,peak_name)

p2
```

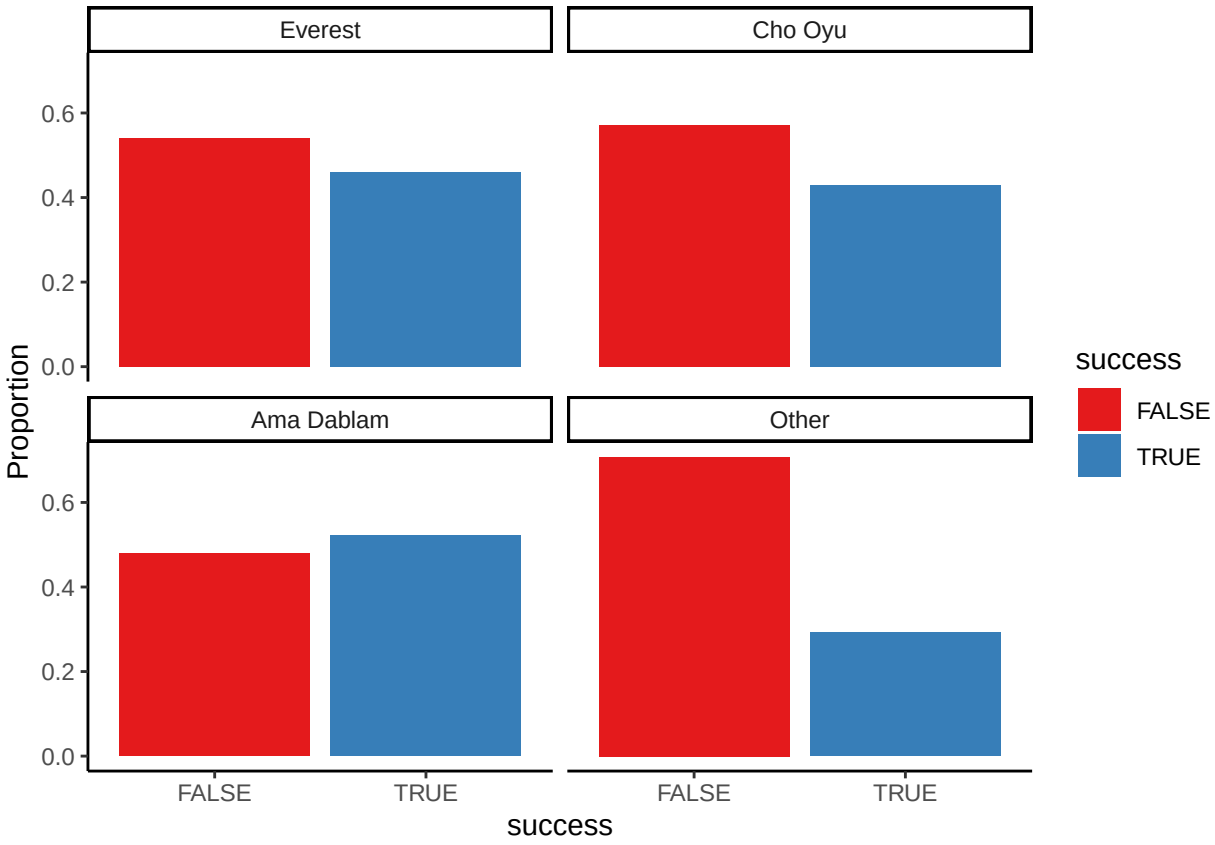
```
## # A tibble: 8 x 3
##   success peak_name      n
##   <lgl>   <fct>      <int>
## 1 FALSE   Everest    11777
## 2 FALSE   Cho Oyu     5068
## 3 FALSE   Ama Dablam  4023
## 4 FALSE   Other      26451
## 5 TRUE    Everest    10036
## 6 TRUE    Cho Oyu     3822
## 7 TRUE    Ama Dablam  4383
## 8 TRUE    Other      10944
```

```
#This code rearranges the data set to have a column for the top three mountains and prints the data tab
table2<-p2%>%
   pivot_wider(names_from = "peak_name", values_from = "n")
table2
```

```
## # A tibble: 2 x 5
##   success Everest `Cho Oyu` `Ama Dablam` Other
##   <lgl>     <int>     <int>        <int> <int>
## 1 FALSE     11777      5068         4023 26451
## 2 TRUE      10036      3822         4383 10944
```

```
# This code creates a new data set with a variable indicating the proportion of successes
coldata<-p2%>%
  group_by(peak_name)%>%
  mutate(prop=n/sum(n))
#This code plots the proportion of successes against each mountain
coldata%>%
  ggplot(aes(x=success, y=prop,fill=success)) +
  geom_col()+
```

```
  facet_wrap(vars(peak_name))+
  theme_classic()+
  ylab("Proportion")+
  scale_fill_brewer(palette = "Set1")
```



Discussion: *Based on the plots, which peak is ascended appears to have a large effect on the proportion of success. We can see this by comparing the blue and red bars in the plot, we see that there is major shifting between each mountain successes. We notably see that for attempts to ascend peaks other than the top three that there are significantly more failures. However we would have to run statistical analyses to determine whether any of these observed differences is statistically significant.*