Project 2

In this project, you will be working with a dataset about the members of Himalayan expeditions:

```
members <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/
tidytuesday/master/data/2020/2020-09-22/members.csv')
```

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/.

Introduction:

The dataset used in the project is about Himalayan expeditions which are taken from the Himalayan Database, a compilation of records for all expeditions that have climbed in the Nepal Himalaya ranging from 1905 to 2019. Each record in the dataset contains information including the name of the mountain (peak_name), the year of the expedition (year), the season (season), the age of the expedition member (age), their citizenship (citizenship), whether they used oxygen (oxygen_used), and whether they successfully summitted the peak (success). The followings are 2 questions I attempted to answer:

Question 1: What are the top 10 countries with the highest number of people who summitted?

For the first question, I need columns success which contains TRUE or FALSE denoting if the individual successfully summitted the top, citizenship denoting different countries the individual is from.

Question 2: How does the percentage of people who summitted the peak and the percentage of death change over time?

For the second question, we need year variable which tells us about the time when the individual goes on a climb, died which contains TRUE or FALSE denoting if the individual has died, and success which contains TRUE or FALSE denoting if the individual successfully summitted the top.

Approach: The approach for the first question is as below: we need to group the members by success and citizenship so we can get the number of individuals from different countries who successfully go to the peak. Then we summarize and calculate the count of successful attempts of people with different nationalities and named the column suc. We arranged the members data set to sort it by descending number of suc. We ungrouped the members so data can be manipulated freely, and we extracted the top 10 rows for the 10 countries with the highest number of successful attempts to the peak by slice.

We plotted the barplot in order to visualize the result. By placing countries on the y axis, we are able to save some space on displaying.

For the second question, we grouped the members by year to prepare the data set for time series analysis. After that, since success and died are boolean variables, we used summary to aggregate the sum of each variable. Then we created column d_rate and s_rate which denote the proportion of indiviudals who went climbing in a specific year that have died or successfully summitted, correspondingly by using mutate. I created a line graph with both success rate and death rate on the y axis, with year being the x variable.

Analysis:

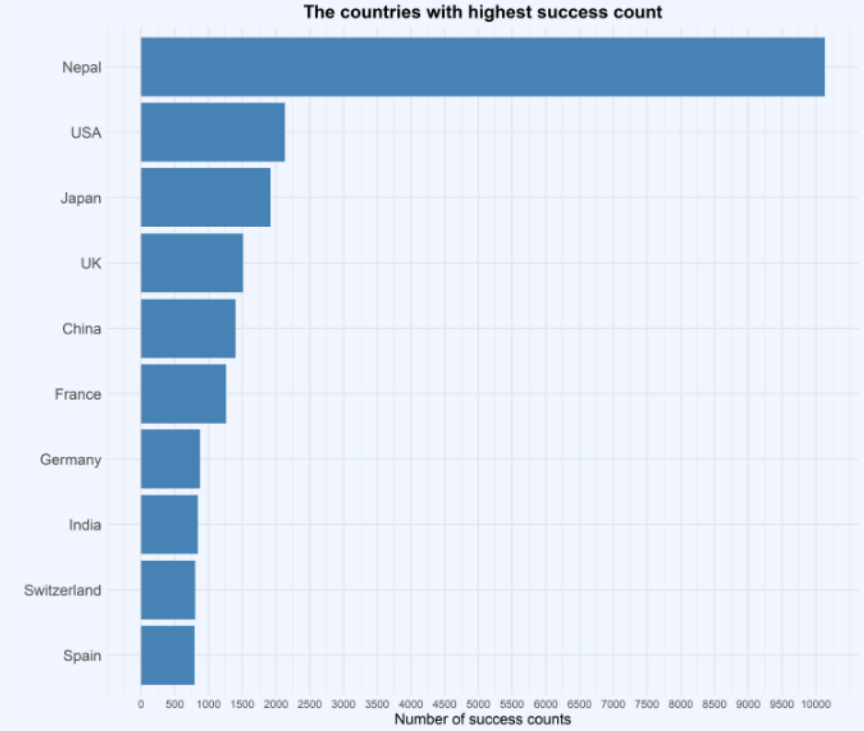
```
# Question 1:

#Prepare dataset for plotting
g<- members |>
  group_by(success, citizenship) |>
  summarize(
    n=n(),
    suc=sum(success=="TRUE", na.rm=TRUE))|>
  arrange(desc(suc)) |>
  ungroup() |>
  slice(1:10) |>
  mutate(citizenship=fct_reorder(citizenship, suc))
```

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

```
#plotting the bar plot
p <- ggplot(g, aes(y = citizenship, x = suc)) +
  geom_col(fill = "steelblue") +
  scale_x_continuous(breaks = seq(0, max(g$suc), by = 500)) +
  scale_y_discrete() +
  labs(
    title = "The countries with highest success count",
    x = "Number of success counts",
    y = ""
  ) +
  theme_minimal(base_size = 20) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 30, face = "bold"),
    axis.title.x = element_text(size = 25),
    axis.title.y = element_text(size = 25),
    axis.text.x = element_text(size = 18),
    axis.text.y = element_text(size = 25),
    panel.background = element_rect(fill = "aliceblue", color = NA),
    legend.box.background = element_rect(fill = "aliceblue", color = NA),
    panel.grid.major = element_line(color = "gray90"),
    plot.background = element_rect(fill = "aliceblue", color = NA),
```

```
    legend.background = element_rect(fill = "aliceblue", color = NA)
  )
p
```



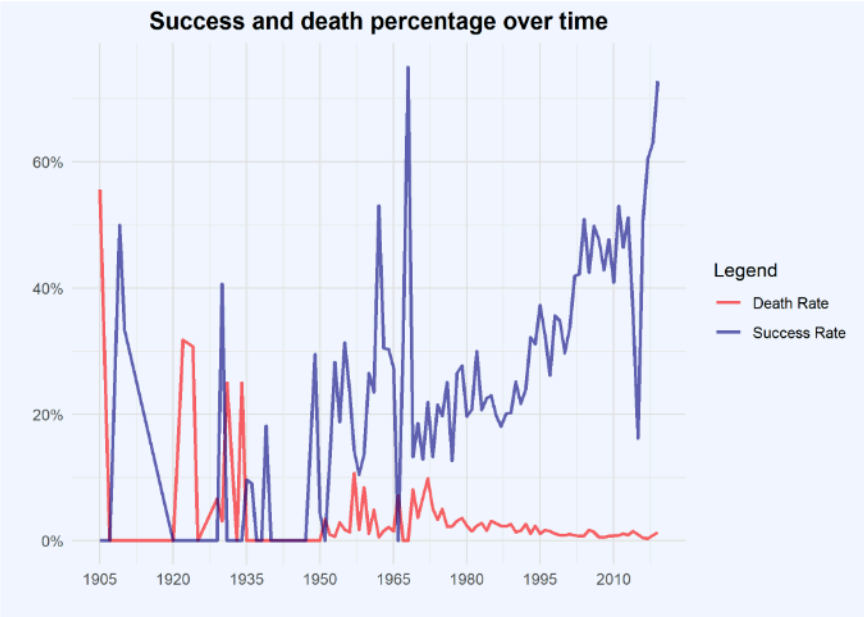

```
# Q2:

#prepare dataset for plotting
g2<- members |>
  group_by(year)|>
  summarize(n=n(),
            die=sum(died=="TRUE", na.rm=TRUE),
            succ=sum(success=="TRUE", na.rm=TRUE))|>
  mutate(d_rate=die*100/n,
         s_rate=succ*100/n)

#Plotting line graph
ggplot(g2)+
  geom_line(aes(x=year, y=d_rate, color="Death Rate"), size=1, alpha=0.6)+
  geom_line(aes(x=year, y=s_rate, color="Success Rate"), size=1, alpha=0.6)+
   scale_color_manual(values = c("Death Rate" = "red", "Success Rate" =
"navyblue"))+
  scale_x_continuous(breaks=seq(min(g2$year), max(g2$year), by=15))+
  scale_y_continuous(labels = scales::percent_format(scale = 1))+
  labs(
```

```
    title="Success and death percentage over time",
    x="",
    y="",
    color="Legend"
  )+
  theme_minimal()+
  theme(
    plot.title= element_text(hjust=0.5, size=14, face="bold"),
        panel.background=element_rect(
          fill="aliceblue",
          color=NA
        ),
        legend.box.background=element_rect(
          fill="aliceblue",
          color=NA
        ),
        panel.grid.major = element_line(color = "gray90"),
        plot.background=element_rect(
          fill="aliceblue",
          color=NA
        ),
        legend.background=element_rect(
          fill="aliceblue",
          color=NA
        ))
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
ℹ Please use `linewidth` instead.
```



Discussion:

According to the first bar graph, the country Nepal has the highest number of successes in total: it has over 10,000 success counts in total, which is almost 5 times of counts of USA, which is the second highest count country. Starting from the USA, the counts are decreasing not as dramatically. Japan, the 3rd place has around 500 less counts than US does. Followed by UK, China, France, Germany, India, Switzerland and Spain.

The second graph, which is a time series plot with 2 colored lines, with red line representing death rate, and blue line representing success rate, shows audience the change of success and death percentage among all mountain climbers over time, ranging from 1905 to 2019. First, when we look at the death rate, we can see the clear trend downward. Around time 1905, the death rate peaked at around 55%. Immediately after, it drops to zero percent and it was constant for around 15 years till when it went up till 30% and stay around 30% for several years. It decreases till zero for 1 year and fluctuates heavily at the range of 0% to 35% for around 10 years. After that, it stayed flat at 0 for 15 years and started have slight fluctuation at the range from 3% to 10% till 2019. The success rate, on the other hand, displays a slight upward trend. At 1905, the success rate went straight to 50% in a few years and dropped heavily to 0%. It stayed at 0 for 10 years and went straight up to 40% and straight down to 0% immediately after. There are 2 more fluctuation from 10-20% to 0% after, and there were 10 years of 0 success rate. Starting from around 1950, where the upward trend starts to display, and a record high occurred in year 1965, where it goes up to more than 75%. Despite the big volatility in the line, the line only touched 0% once and began to steadily increase: from 15% in 1967 to 45% in 2010. There was a big drop from 50% to 17% in year 2018-2019, but it recovers quickly and went back to more than 70%.