Day Four: Data Fundamentals and Intro to RStudio Environment

SDS 192: Introduction to Data Science

Lindsay Poirier Statistical & Data Sciences, Smith College

Spring 2022

```
global_landslide <- read.csv("https://data.nasa.gov/api/views/dd9e-wu2v/rows.csv")
```

1. Identify a unique key in this dataset. Check whether this unique key repeats.

```
# Check whether the unique key you've identified repeats

length(unique(global_landslide$event_id)) == nrow(global_landslide)
```

```
## [1] TRUE
```

2. Calculate the total fatality count in this dataset and total injury count in this dataset. Calculate the percentage of NA entries in each of these variables.

```
# Calculate the total fatality count and total injury count

sum(global_landslide$fatality_count, na.rm = TRUE)
```

```
## [1] 31061
```

```
sum(global_landslide$injury_count, na.rm = TRUE)
```

```
## [1] 4029
```

```
# Calculate the % NA values

sum(is.na(global_landslide$fatality_count)) / length(global_landslide$fatality_count) * 100
```

```
## [1] 12.55325
```

```
sum(is.na(global_landslide$injury_count)) / length(global_landslide$injury_count) * 100
```

```
## [1] 51.42754
```

3. Uncomment and complete the code below to generate a new column with a newspaper headline for each row in the dataset. Your headline should include at least five variables from the dataset, concatenated with narrative text.

```
# Generate a new column with a newspaper headline for each landslide

global_landslide$headline <- paste("According to",
                                   global_landslide$source_name,
                                   "on",
                                   global_landslide$event_date,
                                   "a",
                                   global_landslide$landslide_size,
                                   global_landslide$landslide_category,
                                   "occurred, killing",
                                   global_landslide$fatality_count,
                                   "people",
                                   sep = " ")
```

4. Check the possible values in `landslide_size`. Factor this variable, setting the levels from smallest to largest. Table the unique values in `landslide_size` and `landslide_size-factored`.

```
# Check the possible values in landslide_size

unique(global_landslide$landslide_size)
```

```
## [1] "large"        "small"        "medium"       "unknown"      "very_large"
## [6] ""             "catastrophic"
```

```
# Uncomment below and factor landslide_size
global_landslide$landslide_size_factored <- factor(global_landslide$landslide_size,
                                                   levels = c("small",
                                                              "medium",
                                                              "large",
                                                              "very_large",
                                                              "catastrophic",
                                                              "unknown",
                                                              ""))

# Compare the outputs when you run the table() function with `landslide_size` vs. with global_landslide
table(global_landslide$landslide_size)
```

```
##
##              catastrophic        large       medium        small      unknown
##             9            3          750         6551         2767          851
##   very_large
##           102
```

```
table(global_landslide$landslide_size_factored)
```

```
##
##        small       medium        large   very_large catastrophic      unknown
##         2767         6551          750          102            3          851
##
##            9
```