Exercise: **Data Management**

Day 2, Part A

1. Load the Nigeria health metrics data set ('data/nigeria_healthmap.csv').

```
> main_dir <- "C:/Users/ngraetz/Documents/repos/r_training_penn/"  # CHANGE TO YOUR LOCAL COPY O
> data <- read.csv(paste0(main_dir, "data/nigeria_healthmap.csv"))
```

   a. Keep just the rows where the `geography` variable is 'National'.

```
> data <- data[data$geography == "National", ]
```

   b. Keep just the `year`, `indicator`, `units`, `estimate`, `ci_lb`, and `ci_ub` variables.

```
> data <- data[, c("year", "indicator", "units", "estimate", "ci_lb", "ci_ub")]
```

   c. Load the population counts for Nigeria and merge onto the health metrics data set ('data/nigeria_pop.csv').

```
> pop <- read.csv(paste0(main_dir, "data/nigeria_pop.csv"))
> data <- merge(data, pop, by = "year")
```

   d. Sort the data by year and then indicator.

```
> data <- data[order(data$year, data$indicator), ]
```

   e. Rename the variables `estimate`, `ci_lb`, and `ci_ub` to `est`, `lwr`, and `upr`, respectively.

```
> data <- plyr::rename(data, c(estimate = "est", ci_lb = "lwr", ci_ub = "upr"))
```

   f. Save your formatted data as a .rds file in the 'output' folder of your main directory.

```
> saveRDS(data, file = paste0(main_dir, "/output/nigeria_formatted_data.rds"))
```

2. Load in the Ebola geospatial data (two files: 'data/ebola_point_data.csv', and 'data/ebola_polygon_data.csv').

```
> point <- read.csv(paste0(main_dir, "data/ebola_point_data.csv"))
> poly <- read.csv(paste0(main_dir, "data/ebola_polygon_data.csv"))
```

   a. Combine the point and polygon data into one data frame.

```
> data <- rbind(point, poly)
```

   b. Keep only the `UNIQ_ID`, `Country`, `Virus`, `LAT`, `LONG`, `STR_YEAR`, `OB_CASE`, `OB_DEATH`, and `CASE_TYPE` variables.

```
> data <- data[, c("UNIQ_ID", "Country", "Virus", "LAT", "LONG", "STR_YEAR", "OB_CASE",
+     "OB_DEATH", "CASE_TYPE")]
```

   c. Rename these columns to `ID`, `Country`, `Virus`, `Latitude`, `Longitude`, `Year`, `Cases`, `Deaths`, and `Type`, respectively.

```
> names(data) <- c("ID", "Country", "Virus", "Latitude", "Longitude", "Year", "Cases", "Deat
+     "Type")
```

   d. Keep only rows that refer to 'index' type cases.

```
> data <- data[data$Type == "index", ]
```

   e. Sort the data by country and year.

```
> data <- data[order(data$Country, data$Year), ]
```

f. Save your formatted data as a .csv file in the 'output' folder of your main directory.

```
> write.csv(data, file = paste0(main_dir, "output/ebola_all_index_data.csv"), row.names = F)
```