DS assessment question 1

Jianing Yao

2023-08-26

Web scraping

```
library(rvest)

# URL of the wikipedia page on natural disasters
url <- "https://en.wikipedia.org/wiki/List_of_natural_disasters_by_death_toll"
content <- read_html(url)
tables <- html_table(content)
century20 <- tables[[2]]
head(century20)
```

```
## # A tibble: 6 x 6
##    Year `Death toll` Event                                  Countrie~1 Type  Date
##   <int> <chr>        <chr>                                  <chr>      <chr> <chr>
## 1  1900 6,000-8,000  1900 Galveston hurricane               United St~ Trop~ Sept~
## 2  1901 9,500        1901 eastern United States heat wave   United St~ Heat~ June~
## 3  1902 29,000       1902 eruption of Mount Pelée           Martinique Volc~ Apri~
## 4  1903 3,500        1903 Manzikert earthquake              Turkey     Eart~ Apri~
## 5  1904 400          1904 Sichuan earthquake                China      Eart~ Augu~
## 6  1905 20,000+      1905 Kangra earthquake                 India      Eart~ Apri~
## # ... with abbreviated variable name 1: `Countries affected`
```

```
century21 <- tables[[3]]
head(century21)
```

```
## # A tibble: 6 x 6
##    Year `Death toll` Event                                    Count~1 Type  Date
##   <int> <chr>        <chr>                                    <chr>   <chr> <chr>
## 1  2001 20,005       2001 Gujarat earthquake                  India   Eart~ Janu~
## 2  2002 1,030        2002 Indian heat wave                    India   Heat~ May
## 3  2003 72,000       2003 European heat wave                  Europe  Heat~ July~
## 4  2004 227,898      2004 Indian Ocean earthquake and tsuna~ Indone~ Eart~ Dece~
## 5  2005 87,351       2005 Kashmir earthquake                  India,~ Eart~ Octo~
## 6  2006 5,782        2006 Yogyakarta earthquake               Indone~ Eart~ May ~
## # ... with abbreviated variable name 1: `Countries affected`
```

```
century20 <- as.data.frame(century20)
century21 <- as.data.frame(century21)
```

Merge the 20th and 21st century data frames

```
century20$Century <- rep('20th', nrow(century20))
century21$Century <- rep('21st', nrow(century21))
```

```
disasters <- rbind(century20, century21)
disasters$Century <- as.factor(disasters$Century)
levels(disasters$Century) <- c("20th Century", "21st Century")
disasters$Type <- as.factor(disasters$Type)
disasters$Event <- gsub("\\[\\d+\\]", "", disasters$Event)
disasters$`Death toll` <- gsub("\\[\\d+\\]", "", disasters$`Death toll`)
```

Convert the death toll to numbers

```
disasters$`Death toll` <- gsub("[+,]", "", disasters$`Death toll`)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

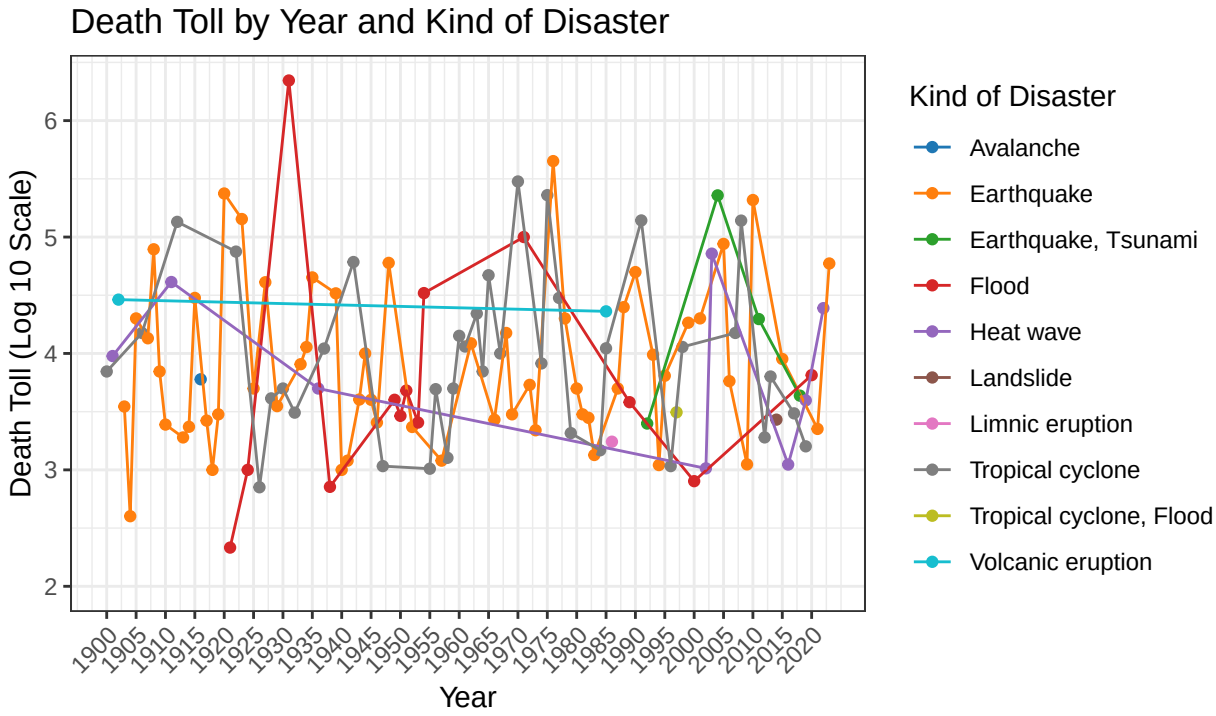
```
convert_range <- function(death_toll){
  if(grepl("-", death_toll)){
    mean(as.numeric(strsplit(death_toll, "-")[[1]]))
  } else{
    as.numeric(death_toll)
  }
}

disasters$`Death toll` <- sapply(disasters$`Death toll`, convert_range)
```

Plot the death toll

```
library(ggplot2)
distinct_colors <- c("#1f77b4", "#ff7f0e", "#2ca02c", "#d62728", "#9467bd",
                     "#8c564b", "#e377c2", "#7f7f7f", "#bcbd22", "#17becf")
plt <- ggplot(disasters, aes(x = Year, y = log10(`Death toll`), color = Type)) +
     geom_point() +
     geom_line() +
     scale_x_continuous(breaks = seq(min(disasters$Year), max(disasters$Year), by = 5)) +
     scale_color_manual(values = distinct_colors) +
     labs(title = "Death Toll by Year and Kind of Disaster",
          x = "Year",
          y = "Death Toll (Log 10 Scale)",
          color = "Kind of Disaster") +
     theme_bw() +
     expand_limits(y = 2) +
     theme(
         aspect.ratio = 0.7,
         axis.text.x = element_text(angle = 45, hjust = 1)
    )
ggsave("plot.png", plt, width = 10, height = 6, units = "in", dpi = 300)
```

```
plt
```


Death Toll by Year and Kind of Disaster
Death Toll (Log 10 Scale)
6
5
4
3
2
1900
1905
1910
1915
1920
1925
1930
1935
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year
Kind of Disaster
Avalanche
Earthquake
Earthquake, Tsunami
Flood
Heat wave
Landslide
Limnic eruption
Tropical cyclone
Tropical cyclone, Flood
Volcanic eruption