Assignment2

Haoyi Wei

2022-12-18

Explore the NOAA Storm Database

Synopsis

The basic goal of this assignment is to expore the NOAA Storm Database and answer some basic questions about severe weather events. The data analyssi aims to addres two questions: (1) Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health? (2) Across the US,, which types of events have the greatest economic consequences?

Loading and Processing the Raw Data

From the Coursera website, we obtained the storm dataset.

Read the data

```
library(readr)
storm_data <- read_csv("repdata-data-StormData.csv")
```

```
## Rows: 902297 Columns: 37
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (18): BGN_DATE, BGN_TIME, TIME_ZONE, COUNTYNAME, STATE, EVTYPE, BGN_AZI,...
## dbl (18): STATE__, COUNTY, BGN_RANGE, COUNTY_END, END_RANGE, LENGTH, WIDTH, ...
## lgl  (1): COUNTYENDN
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

After reading in the data, we check the first few rows in the datasets

```
dim(storm_data)
```

```
## [1] 902297     37
```

```
head(storm_data)
```

```
## # A tibble: 6 × 37
##   STATE__ BGN_DATE   BGN_T…¹ TIME_…² COUNTY COUNT…³ STATE EVTYPE BGN_R…⁴ BGN_AZI
##     <dbl> <chr>      <chr>   <chr>     <dbl> <chr>   <chr> <chr>   <dbl> <chr>
## 1       1 4/18/1950… 0130    CST          97 MOBILE  AL    TORNA…      0 <NA>
## 2       1 4/18/1950… 0145    CST           3 BALDWIN AL    TORNA…      0 <NA>
## 3       1 2/20/1951… 1600    CST          57 FAYETTE AL    TORNA…      0 <NA>
## 4       1 6/8/1951 … 0900    CST          89 MADISON AL    TORNA…      0 <NA>
## 5       1 11/15/195… 1500    CST          43 CULLMAN AL    TORNA…      0 <NA>
## 6       1 11/15/195… 2000    CST          77 LAUDER… AL    TORNA…      0 <NA>
## # … with 27 more variables: BGN_LOCATI <chr>, END_DATE <chr>, END_TIME <chr>,
## #   COUNTY_END <dbl>, COUNTYENDN <lgl>, END_RANGE <dbl>, END_AZI <chr>,
## #   END_LOCATI <chr>, LENGTH <dbl>, WIDTH <dbl>, F <dbl>, MAG <dbl>,
## #   FATALITIES <dbl>, INJURIES <dbl>, PROPDMG <dbl>, PROPDMGEXP <chr>,
## #   CROPDMG <dbl>, CROPDMGEXP <chr>, WFO <chr>, STATEOFFIC <chr>,
## #   ZONENAMES <chr>, LATITUDE <dbl>, LONGITUDE <dbl>, LATITUDE_E <dbl>,
## #   LONGITUDE_ <dbl>, REMARKS <chr>, REFNUM <dbl>, and abbreviated variable …
## # ℹ Use `colnames()` to see all variable names
```

clean the data

```
tidy_storm <- storm_data[,c('EVTYPE','FATALITIES','INJURIES', 'PROPDMG', 'PROPDMGEXP', 'CROPDMG', 'CROPDMGEXP')]
head(tidy_storm)
```

```
## # A tibble: 6 × 7
##   EVTYPE  FATALITIES INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP
##   <chr>        <dbl>    <dbl>   <dbl> <chr>        <dbl> <chr>
## 1 TORNADO          0       15    25   K                0 <NA>
## 2 TORNADO          0        0     2.5 K                0 <NA>
## 3 TORNADO          0        2    25   K                0 <NA>
## 4 TORNADO          0        2     2.5 K                0 <NA>
## 5 TORNADO          0        2     2.5 K                0 <NA>
## 6 TORNADO          0        6     2.5 K                0 <NA>
```

```
str(tidy_storm)
```

```
## tibble [902,297 × 7] (S3: tbl_df/tbl/data.frame)
##  $ EVTYPE    : chr [1:902297] "TORNADO" "TORNADO" "TORNADO" "TORNADO" ...
##  $ FATALITIES: num [1:902297] 0 0 0 0 0 0 0 0 0 1 0 ...
##  $ INJURIES  : num [1:902297] 15 0 2 2 2 6 1 0 14 0 ...
##  $ PROPDMG   : num [1:902297] 25 2.5 25 2.5 2.5 2.5 2.5 2.5 25 25 ...
##  $ PROPDMGEXP: chr [1:902297] "K" "K" "K" "K" ...
##  $ CROPDMG   : num [1:902297] 0 0 0 0 0 0 0 0 0 0 ...
##  $ CROPDMGEXP: chr [1:902297] NA NA NA NA ...
```

Results

Question 1

Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health?

```
# check number of missing values
sum(is.na(tidy_storm[,c('FATALITIES','INJURIES')]))
```

```
## [1] 0
```

```
# make the health demage dataset
library(dplyr)
```

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

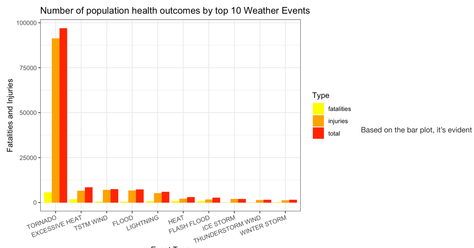
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
library(tidyr)
tidy_q1 <- tidy_storm %>%
          group_by(EVTYPE) %>%
          summarize(fatalities = sum(FATALITIES), injuries= sum(INJURIES), .groups='drop') %>%
          mutate(total = fatalities + injuries) %>%
          select(EVTYPE, total,fatalities, injuries) %>%
          arrange(-total,-fatalities,-injuries) %>%
          slice(1:10) %>% # select top ten events
          mutate(evtype=factor(EVTYPE, levels=EVTYPE)) %>%
          gather(key = Type, value = Value,total,fatalities,injuries)
```

```
# make the plot
library(ggplot2)

ggplot(tidy_q1, aes(evtype, Value, fill=Type)) +
          geom_bar(position="dodge",stat="identity") +
          theme_bw() +
          theme(axis.text.x= element_text(angle=20, vjust=1, hjust=1)) +
          xlab("Event Type") +
          ylab("Fatalities and Injuries") +
          ggtitle("Number of population health outcomes by top 10 Weather Events") +
          scale_fill_manual(values=c("yellow","orange","red"))
```



Based on the bar plot, it's evident that tornadoes have the highest impact on the population health, since it causes the most fatalities and injuries.

Question 2

```
unique(storm_data$PROPDMGEXP)
```

```
##  [1] "K" "M" NA  "B" "m" "+" "0" "5" "6" "?" "4" "2" "3" "h" "7" "H" "-" "1" "8"
```

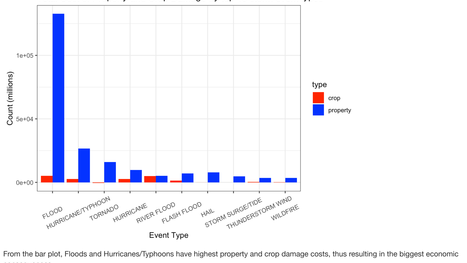
```
unique(storm_data$CROPDMGEXP)
```

```
## [1] NA  "M" "K" "m" "B" "?" "0" "k" "2"
```

```
cost <- function(x) {
  if (x == "H")
    1E-4
  else if (x == "K")
    1E-3
  else if (x == "M")
    1
  else if (x == "B")
    1E3
  else
    1-6
}
```

```
economic <- tidy_storm %>%
          filter(is.na(CROPDMGEXP)!=1 & is.na(PROPDMGEXP)!=1) %>%
      mutate(prop_dmg = PROPDMG*sapply(PROPDMGEXP, FUN = cost),
             crop_dmg = CROPDMG*sapply(CROPDMGEXP, FUN = cost), .keep="unused") %>%
      group_by(EVTYPE) %>%
      summarise(property = sum(prop_dmg), crop = sum(crop_dmg), .groups='drop') %>%
      arrange(desc(property), desc(crop)) %>%
      slice(1:10) %>%
      gather(key = type, value = value, property, crop)

ggplot(data=economic, aes(reorder(EVTYPE, -value), value, fill=type)) +
  geom_bar(position = "dodge", stat='identity') +
  labs(x="Event Type", y="Count (millions)") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 25, vjust=0.5)) +
  ggtitle("Total Cost of Property and Crop Damage by top 10 storm event types") +
  scale_fill_manual(values=c("red", "blue"))
```



From the bar plot, Floods and Hurricanes/Typhoons have highest property and crop damage costs, thus resulting in the biggest economic consequences.

Conclusion

Based on the analysis, resources should be directed towards dealing with tornadoes for the safety and health of population by building better infrastructure or early warning systems. As for dealing with hurricanes and typhoons, there should be more funding for innovation in developing better systems and infrastructure to safeguard these properties and crops to prevent damages as much as possible.

Reference

Roger D. Peng's Example

Benedict Neo Yao En's Example