NOAA Storm Database Analysis

Isaac G Veras

05/10/2023

Introduction

Storms and other severe weather events can cause both public health and economic problems for communities and municipalities. Many severe events can result in fatalities, injuries, and property damage, and preventing such outcomes to the extent possible is a key concern.

This project involves exploring the **U.S. National Oceanic and Atmospheric Administration's (NOAA)** storm database. This database tracks characteristics of major storms and weather events in the United States, including when and where they occur, as well as estimates of any fatalities, injuries, and property damage.

Questions

The analysis is trying to answer the following 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 United States, which types of events have the greatest economic consequences?

Package installation:

```
if (!require("pacman")) install.packages("pacman")
```

```
## Carregando pacotes exigidos: pacman
```

```
pacman::p_load(pacman,     # Package Manager
               knitr,      # Transform R Markdown documents into various output formats
               plyr,       # Data manipulation
               data.table, # Manipulate, process and analyze large data sets
               tidyverse   # Data organization
)
```

1. Data Processing:

The data for this assignment come in the form of a comma-separated-value file compressed via the bzip2 algorithm to reduce its size.

```
data_url  <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
file_name <- "StormData.csv.bz2"

if (!file.exists(file_name)) {
    download.file(data_url, file_name, method = "curl")
}
```

1.1 Load Data into dataset

```
storm_data <- read.csv(file_name,
                       header = TRUE,
                       sep    = ","
)
```

The raw data structure consist of 902297 row(s) and 37 columns

List of available columns:

```
##  [1] "STATE__"    "BGN_DATE"   "BGN_TIME"   "TIME_ZONE"  "COUNTY"
##  [6] "COUNTYNAME" "STATE"      "EVTYPE"     "BGN_RANGE"  "BGN_AZI"
## [11] "BGN_LOCATI" "END_DATE"   "END_TIME"   "COUNTY_END" "COUNTYENDN"
## [16] "END_RANGE"  "END_AZI"    "END_LOCATI" "LENGTH"     "WIDTH"
## [21] "F"          "MAG"        "FATALITIES" "INJURIES"   "PROPDMG"
## [26] "PROPDMGEXP" "CROPDMG"    "CROPDMGEXP" "WFO"        "STATEOFFIC"
## [31] "ZONENAMES"  "LATITUDE"   "LONGITUDE"  "LATITUDE_E" "LONGITUDE_"
## [36] "REMARKS"    "REFNUM"
```

Check first five rows of raw data:

```
head(storm_data)
```

```
##   STATE__           BGN_DATE BGN_TIME TIME_ZONE COUNTY COUNTYNAME STATE  EVTYPE
## 1       1  4/18/1950 0:00:00     0130       CST     97     MOBILE    AL TORNADO
## 2       1  4/18/1950 0:00:00     0145       CST      3    BALDWIN    AL TORNADO
## 3       1  2/20/1951 0:00:00     1600       CST     57    FAYETTE    AL TORNADO
## 4       1   6/8/1951 0:00:00     0900       CST     89    MADISON    AL TORNADO
## 5       1 11/15/1951 0:00:00     1500       CST     43    CULLMAN    AL TORNADO
## 6       1 11/15/1951 0:00:00     2000       CST     77 LAUDERDALE    AL TORNADO
##   BGN_RANGE BGN_AZI BGN_LOCATI END_DATE END_TIME COUNTY_END COUNTYENDN
## 1         0                                               0         NA
## 2         0                                               0         NA
## 3         0                                               0         NA
## 4         0                                               0         NA
## 5         0                                               0         NA
## 6         0                                               0         NA
##   END_RANGE END_AZI END_LOCATI LENGTH WIDTH F MAG FATALITIES INJURIES PROPDMG
## 1         0                      14.0   100 3   0          0       15    25.0
## 2         0                       2.0   150 2   0          0        0     2.5
## 3         0                       0.1   123 2   0          0        2    25.0
## 4         0                       0.0   100 2   0          0        2     2.5
## 5         0                       0.0   150 2   0          0        2     2.5
## 6         0                       1.5   177 2   0          0        6     2.5
##   PROPDMGEXP CROPDMG CROPDMGEXP WFO STATEOFFIC ZONENAMES LATITUDE LONGITUDE
## 1          K       0                                         3040      8812
## 2          K       0                                         3042      8755
## 3          K       0                                         3340      8742
## 4          K       0                                         3458      8626
## 5          K       0                                         3412      8642
## 6          K       0                                         3450      8748
##   LATITUDE_E LONGITUDE_ REMARKS REFNUM
## 1       3051       8806              1
## 2          0          0              2
## 3          0          0              3
## 4          0          0              4
## 5          0          0              5
## 6          0          0              6
```

2. Data Cleansing

For this analysis, only a few columns which are required to answer both questions. So, we create a subset from raw dataset which contains the meaningful variable for this research. The required column are:–

No.	Column	Description
1.	EVTYPE	Type of event recorded
2.	FATALITIES	Number of fatalities reported
3.	INJURIES	Number of people injured reported
4.	PROPDMG	Property damage measurement
5.	PROPDMGEXP	The exponential for Property Damage
6.	CROPDMG	Crop damage measurement
7.	CROPDMGEXP	The exponential for Crop Damage

```
storm_data_select <- select(storm_data, EVTYPE, FATALITIES, INJURIES, PROPDMG, PROPDMGEXP, CROP
DMG, CROPDMGEXP)
```

Check first five rows from subset dataset:

```
head(storm_data_select)
```

```
##    EVTYPE FATALITIES INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP
## 1 TORNADO          0       15    25.0          K       0
## 2 TORNADO          0        0     2.5          K       0
## 3 TORNADO          0        2    25.0          K       0
## 4 TORNADO          0        2     2.5          K       0
## 5 TORNADO          0        2     2.5          K       0
## 6 TORNADO          0        6     2.5          K       0
```

To get the right value, we must change the property damage and crop damage to it’s actual value. The exponential is describe as shown in the table below:–

No.	EXP	Description
1	H	Hundred (10^2)
2	K	Thousand (10^3)
3	M	Million (10^6)
4	B	Billion (10^9)

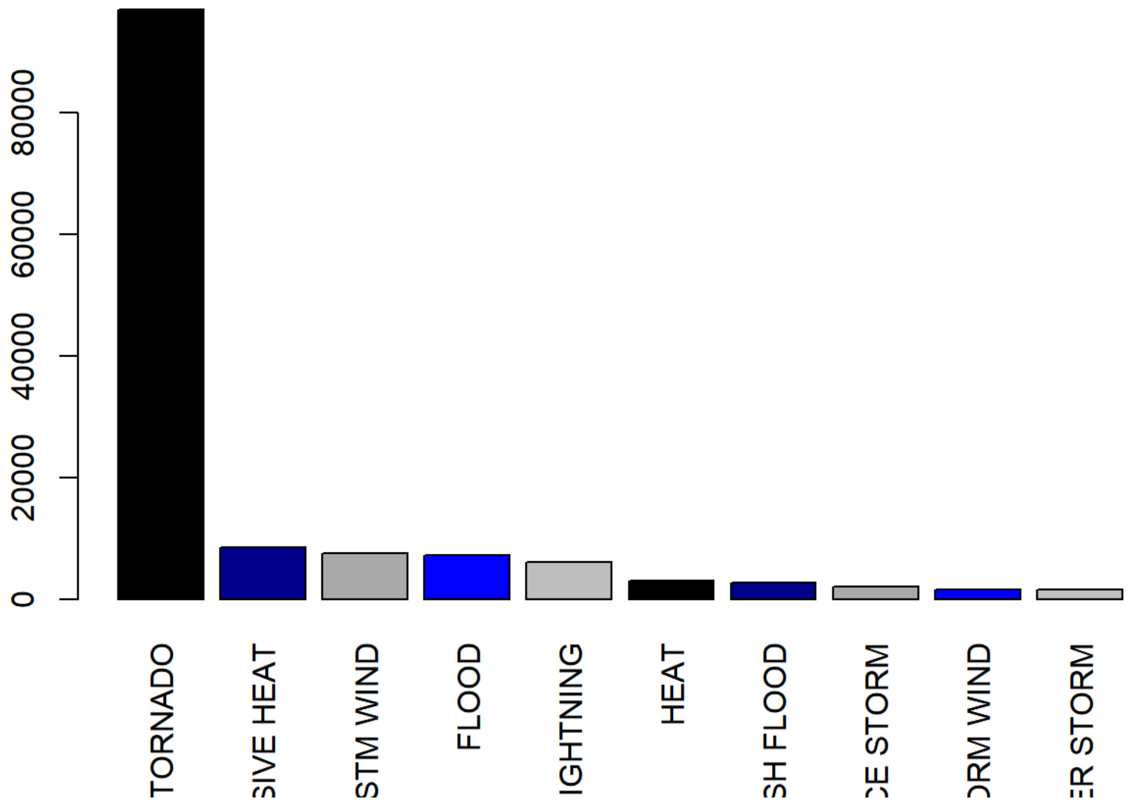
3. Analysing Data

3.1 Events are most harmful with respect to population health.

```
health_effects <- storm_data_select %>%
        group_by(EVTYPE) %>%
        summarise(health_affected = sum(FATALITIES + INJURIES)) %>%
        arrange(desc(health_affected)
        )

top10 <- health_effects[1:10,]

with(top10, barplot(height    = health_affected,
                    names.arg = EVTYPE,
                    las       = 3,
                    col = c("black",
                            "darkblue",
                            "darkgray",
                            "blue",
                            "gray"))
)
```

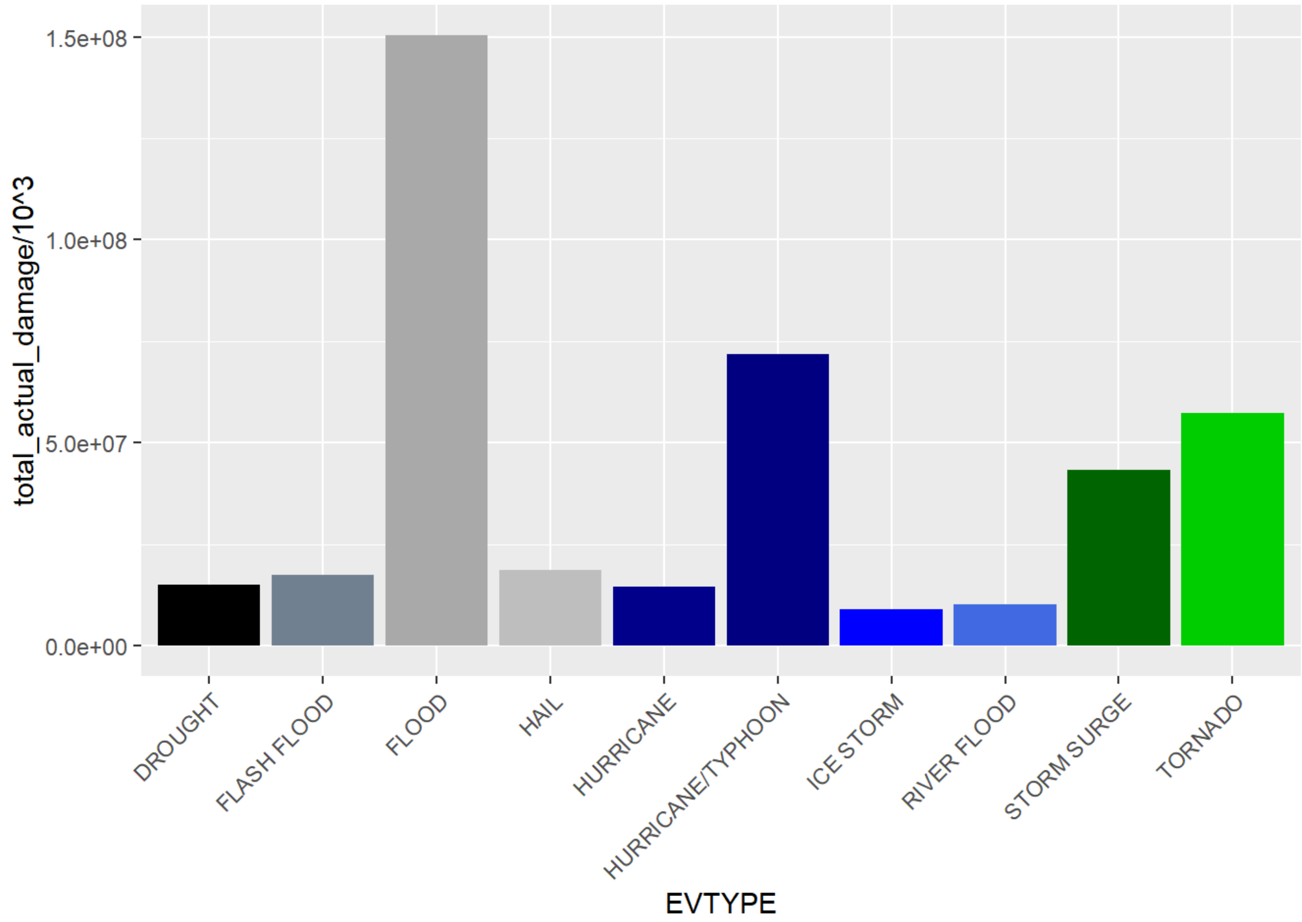


3.2 Events have the greatest economic consequences:

```
economic_effects <- storm_data_select %>%
        mutate(actual_propdmg = case_when(.$PROPDMGEXP == "H" ~ .$PROPDMG * 10^2,
                                          .$PROPDMGEXP == "K" ~ .$PROPDMG * 10^3,
                                          .$PROPDMGEXP == "M" ~ .$PROPDMG * 10^6,
                                          .$PROPDMGEXP == "B" ~ .$PROPDMG * 10^9,
                                          TRUE ~ .$PROPDMG)) %>%
        mutate(actual_cropdmg = case_when(.$CROPDMGEXP == "H" ~ .$CROPDMG * 10^2,
                                          .$CROPDMGEXP == "K" ~ .$CROPDMG * 10^3,
                                          .$CROPDMGEXP == "M" ~ .$CROPDMG * 10^6,
                                          .$CROPDMGEXP == "B" ~ .$CROPDMG * 10^9,
                                          TRUE ~ .$CROPDMG)) %>%
        group_by(EVTYPE) %>%
        summarise(total_actual_damage = sum(actual_propdmg + actual_cropdmg)) %>%
        arrange(desc(total_actual_damage))

clrs <- c("black", "slategray", "darkgray", "gray", "darkblue", "navy", "blue", "royalblue", "d
arkgreen", "green3")
top10 <- economic_effects[1:10,]
ggplot(data = top10, aes(EVTYPE, total_actual_damage / 10^3, fill = EVTYPE)) +
        geom_bar(stat = "identity") +
        guides(fill = FALSE) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
        scale_fill_manual(values = clrs)
```

```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Results:

The chart illustrates the top ten types of events responsible for the most extensive property and crop damage in the United States. Tornadoes emerge as the primary cause of property damage in the United States. Subsequently, floods and flash floods rank second, followed by wind and thunderstorms. Hail stands out as the leading contributor to crop damage in the United States. It is followed by floods and flash floods in the second position, with wind and thunderstorms trailing behind.