Exploring the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database.

12/08/2017

Synopsis

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.

The NOAA storm 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.

The basic goal of this project is to explore the NOAA Storm Database and answer some basic questions about severe weather events.

Data

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

One may download the data at the following link:

Storm Data

There is also some documentation of the database available:

- National Weather Service Storm Data Documentation
- National Climatic Data Center Storm Events FAQ

The events in the database start in the year 1950 and end in November 2011. In the earlier years of the database, there are generally fewer events recorded, most likely due to a lack of good records. More recent years should be considered more complete.

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?

Data Processing

The following analysis starts from loading the raw CSV file containing the data into the variable called *data*. For this analysis, we're going to use **dplyr** and **ggplot2** R packages.

```r
library(dplyr)
library(ggplot2)

data <- read.csv("data.bz2", header = TRUE, stringsAsFactors = FALSE)
```

```r
## Row names
names(data)
```

```
##  [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"
```

For this analysis, we're going to explore the following columns:

Population health
- FATALITIES
- INJURIES

Economic consequences
- PROPDMG
- PROPDMGEXP
- CROPDMG
- CROPDMGEXP

Let's group the data by event type. For each type of event, we'll calculate the total amount of people affected throughout 1950-2011 (separately for fatalities and injuries).

```r
library(dplyr)

d <- tbl_df(data)

## group the data by event type
by_event <- group_by(d, EVTYPE)

## total fatalities for each type of event
fatalities <- summarise(by_event, sum(FATALITIES))

## total injuries for each type of event
injuries <- summarise(by_event, sum(INJURIES))
```

The strategy for calculating economic consequences is the following:

- calculate property damage (PROPDMG) and crop damage (CROPDMG) considering the information from PROPDMGEXP and CROPDMGEXP (multipliers, i.e. M - millions, K - thousands, B - billions)
- sum property damage and crop damage (USD) for each type of event throughout 1950-2011

Let's take a look at property and crop damage multipliers in the raw data:

```r
## property damage unique multipliers
dmg_mult <- sort(unique(tolower(d$PROPDMGEXP)))
print(dmg_mult)
```

```
##  [1] ""  "-" "?" "+" "0" "1" "2" "3" "4" "5" "6" "7" "8" "b" "h" "k" "m"
```

```r
## crop damage unique multipliers
crop_mult <- sort(unique(tolower(d$CROPDMGEXP)))
print(crop_mult)
```

```
## [1] ""  "?" "0" "2" "b" "k" "m"
```

The values in these columns represent exponential multipliers for values in the corresponding data rows. For convenience, we'll create a data frame with these multipliers and numerical values for further calculations. We'll also define a helper function which will help us get numerical values for economic consequences.

```
## ""  "-" "?" "+" "0" "1" "2" "3" "4" "5" "6" "7" "8" "b" "h" "k" "m"
mult <- tibble(key = dmg_mult,
               value = c(1, 0, 0, 0, 1, 10, 100, 1000, 10000,
                         1e+05, 1e+06, 1e+07, 1e+08, 1e+09, 100, 1000, 1e+06))
```

Define a helper function which will help us convert values in PROPDMGEXP and CROPDMGEXP.

```
## helper function
convert_dmg <- function(val, e) {
    m <- filter(mult, key == tolower(e))$value
    val * as.numeric(m)
}
```

Finally, let's add two helper rows to the original data: **prop_dmg** and **crop_dmg**. They will contain USD value of economic consequences for each type of event.

```
d <- mutate(d, prop_dmg = 0)
d <- mutate(d, crop_dmg = 0)

## convert to USD values using our helper function
# for (i in 1:nrow(d)) {
#     d$prop_dmg[i] <- convert_dmg(d$PROPDMG[i], d$PROPDMGEXP[i])
#     d$crop_dmg[i] <- convert_dmg(d$CROPDMG[i], d$CROPDMGEXP[i])
# }
#
# d <- mutate(d, total_ec_dmg = prop_dmg + crop_dmg)
```

Calculate the sum of property damage and crop damage (USD) for each type of event throughout 1950-2011 for each type of event.

```
# ec <- summarise(group_by(d, EVTYPE), sum(total_ec_dmg))
```

Summaries:

```
summary(fatalities)
```

```
##     EVTYPE           sum(FATALITIES)
##  Length:985         Min.   :   0.00
##  Class :character   1st Qu.:   0.00
##  Mode  :character   Median :   0.00
##                     Mean   :  15.38
##                     3rd Qu.:   0.00
##                     Max.   :5633.00
```

```
mean_fatal <- mean(fatalities$`sum(FATALITIES)`)
print(mean_fatal)
```

```
## [1] 15.37563
```

```
summary(injuries)
```

```
##     EVTYPE           sum(INJURIES)
##  Length:985         Min.   :    0.0
##  Class :character   1st Qu.:    0.0
##  Mode  :character   Median :    0.0
##                        Mean   :  142.7
##                        3rd Qu.:    0.0
##                        Max.   :91346.0
```

```
mean_inj <- mean(injuries$`sum(INJURIES)`)
print(mean_inj)
```

```
## [1] 142.668
```

```
# summary(ec)
# mean_ec <- mean(ec$`sum(total_ec_dmg)`)
# print(mean_ec)

fatal <- filter(fatalities, `sum(FATALITIES)` > mean_fatal)
inj <- filter(injuries, `sum(INJURIES)` > mean_inj)


names(fatal) <- c("evt", "ppl")
names(inj) <- c("evt", "ppl")
# names(e) <- c("evt", "usd")

fatal$evt <- factor(fatal$evt, levels = fatal$evt[order(fatal$ppl, decreasing = T)])
inj$evt <- factor(inj$evt, levels = inj$evt[order(inj$ppl, decreasing = T)])
# e$evt <- factor(e$evt, levels = e$evt[order(e$usd, decreasing = T)])
```

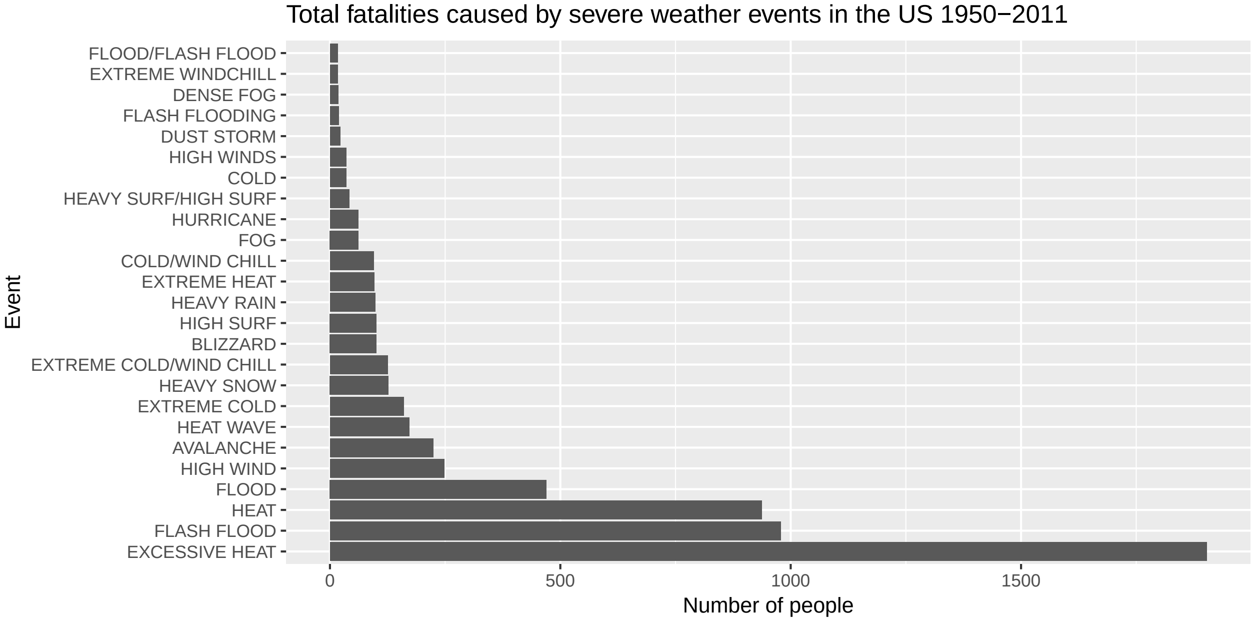
We'll use the above mean values during the data visualization.

Results

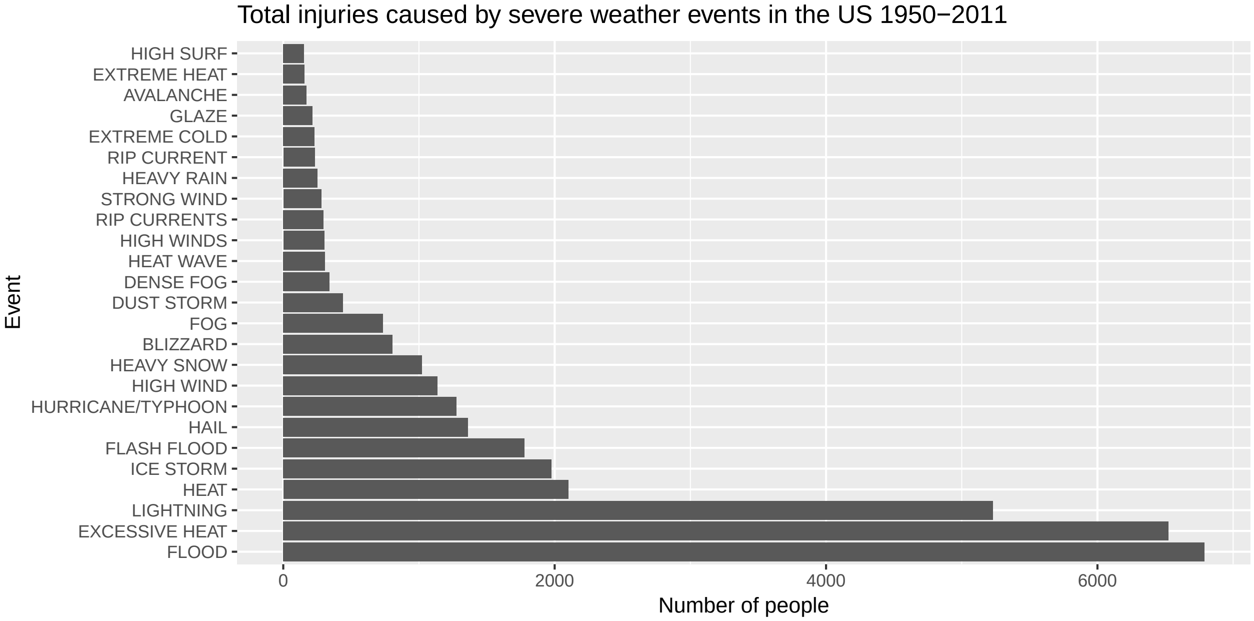
Across the United States, which types of events are most harmful with respect to population health?

```
library(ggplot2)

ggplot(fatal[1:25,], aes(x=evt, y=ppl)) + geom_bar(stat = "identity") +
    coord_flip() + xlab("Event") + ylab("Number of people") +
    ggtitle("Total fatalities caused by severe weather events in the US 1950-2011")
```



```r
ggplot(inj[1:25,], aes(x=evt, y=ppl)) + geom_bar(stat = "identity") + coord_flip() +
    xlab("Event") + ylab("Number of people") +
    ggtitle("Total injuries caused by severe weather events in the US 1950-2011")
```



As we can see from the graphs above, throughout the 1951-2011 most **deaths** caused by severe weather events are:

- *Excessive heat*
- *Flash flood*
- *Heat*
- *Flood*

- *High wind*

Most **injuries** caused by severe weather events are:

- *Flood*
- *Excessive heat*
- *Lightning*
- *Heat*
- *Ice storm*

Across the United States, which types of events have the greatest economic consequences?

```
# ggplot(e[1:25,], aes(x=evt, y=usd)) + geom_bar(stat = "identity") + coord_flip() +
#     xlab("Event") + ylab("USD") +
#     ggtitle("Economic consequences caused by severe weather events in the US 1950-2011")
```

Top 5 types of events have the greatest economic consequences:

- *Flood*
- *Hurricane/Typhoon*
- *Tornado*
- *Storm Surge*
- *Hail*