An analysis of the harmfullness and economic consequences of storms in North-America between 1950 - 2011

Arjen Hunter

24 January 2019

Synopsis

The basic goal of this assignment is to explore the NOAA Storm Database and answer some basic questions about severe weather events. This assignment will use the database to answer the questions below and show the code for your entire analysis. The data analysis will address the following questions:

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

This analysis is being performed as part of the 4th weeks of the Reproducible Research module of the Coursera specialisation course on Data Science offered by John Hopkins University. The data is available online in the form of a comma-separated-value file compressed via the bzip2 algorithm to reduce its size. You can download the file from the course web site.

```
setwd("C:/Users/User/Documents/coursera/Reproducible Research")
# Dowloading data if it's not already done
if(!file.exists("stormData.csv.bz2")) {
  download.file("https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2",
  destfile = "stormData.csv.bz2", method = "curl")
}

# Loading data
df <- read.csv(bzfile("stormData.csv.bz2"), sep=",", header=T)
```

Data Processing

This chapter describes (in words and code) how the data were loaded into R and processed for analysis. In particular, our analysis will start from the raw CSV file containing the data. No preprocessing has been done outside this document.

Firstly we find out which variables are required to answer the two questions above. We'll at least need the event type, all variables describing population health and damage. Reading the documentation that is supplied with the data (National Weather Service Storm Data Documentation) we find the following variables are of interest.

The event type: - EVTYPE Variables describing population health
- FATALITIES
- INJURIES
variables describing economic consequences
- PROPDMG
- PROPDMGEXP
- CROPDMG
- CROPDMGEXP

```
tidydf <- df[,c('EVTYPE','FATALITIES','INJURIES', 'PROPDMG', 'PROPDMGEXP', 'CROPDMG', 'CROPDMGE
XP')]
```

Now that we have combined all the relevant data into one dataframe we can aggregate the data to find out which event is responsible for the most fatalities and sort it in reversed order.

```
fatalities <- aggregate(FATALITIES ~ EVTYPE, data=tidydf, sum)
fatalities <- arrange(fatalities,-fatalities$FATALITIES)
fatalities[1,]
```

```
##    EVTYPE FATALITIES
## 1 TORNADO       5633
```

So tornado's are responsible for 5633 fatalities over the past 61 years

And next to aggregate which event is responsible for the most injuries and sort it in reversed order.

```
injuries <- aggregate(INJURIES ~ EVTYPE, data=tidydf, sum)
injuries <- arrange(injuries,-injuries$INJURIES)
injuries[1,]
```

```
##    EVTYPE INJURIES
## 1 TORNADO    91346
```

And once again it's tornado's that are responsible for 91346 injuries.

Similar to the aggregation of the injuries and fatalities we can see which event is responsible for the greatest economic consequences. The greatest difference being that the data is coded. On page 12 of the documentation is says "Estimates should be rounded to three significant digits, followed by an alphabetical character signifying the magnitude of the number, i.e., 1.55B for $1,550,000,000. Alphabetical characters used to signify magnitude include"K" for thousands, "M" for millions, and "B" for billions."

This means that we can add a column in which we multiply the propdmg and cropdmg with the appropriate multiplier stored in propdmgexp and cropdmgexp, respectively.

```
tidydf <- mutate(tidydf,prop=tidydf$PROPDMG)
tidydf <- mutate(tidydf,crop=tidydf$CROPDMG)
tidydf$crop[tidydf$CROPDMGEXP=="K"] <- tidydf$crop[tidydf$CROPDMGEXP=="K"]*1000
tidydf$crop[tidydf$CROPDMGEXP=="M"] <- tidydf$crop[tidydf$CROPDMGEXP=="M"]*1000000
tidydf$crop[tidydf$CROPDMGEXP=="B"] <- tidydf$crop[tidydf$CROPDMGEXP=="B"]*1000000000
tidydf$prop[tidydf$PROPDMGEXP=="K"] <- tidydf$prop[tidydf$PROPDMGEXP=="K"]*1000
tidydf$prop[tidydf$PROPDMGEXP=="M"] <- tidydf$prop[tidydf$PROPDMGEXP=="M"]*1000000
tidydf$prop[tidydf$PROPDMGEXP=="B"] <- tidydf$prop[tidydf$PROPDMGEXP=="B"]*1000000000
```

Finally we get to aggregate the damage that was caused to crop and property by the type of event and sorting it in reversed order.

```
damage <- aggregate(crop + prop ~ EVTYPE, data=tidydf, sum)
damage <- arrange(damage,-damage$`crop + prop`)
damage[1:5,]
```

```
##              EVTYPE  crop + prop
## 1             FLOOD 150319678257
## 2 HURRICANE/TYPHOON  71913712800
## 3           TORNADO  57340614060
## 4       STORM SURGE  43323541000
## 5              HAIL  18752904943
```

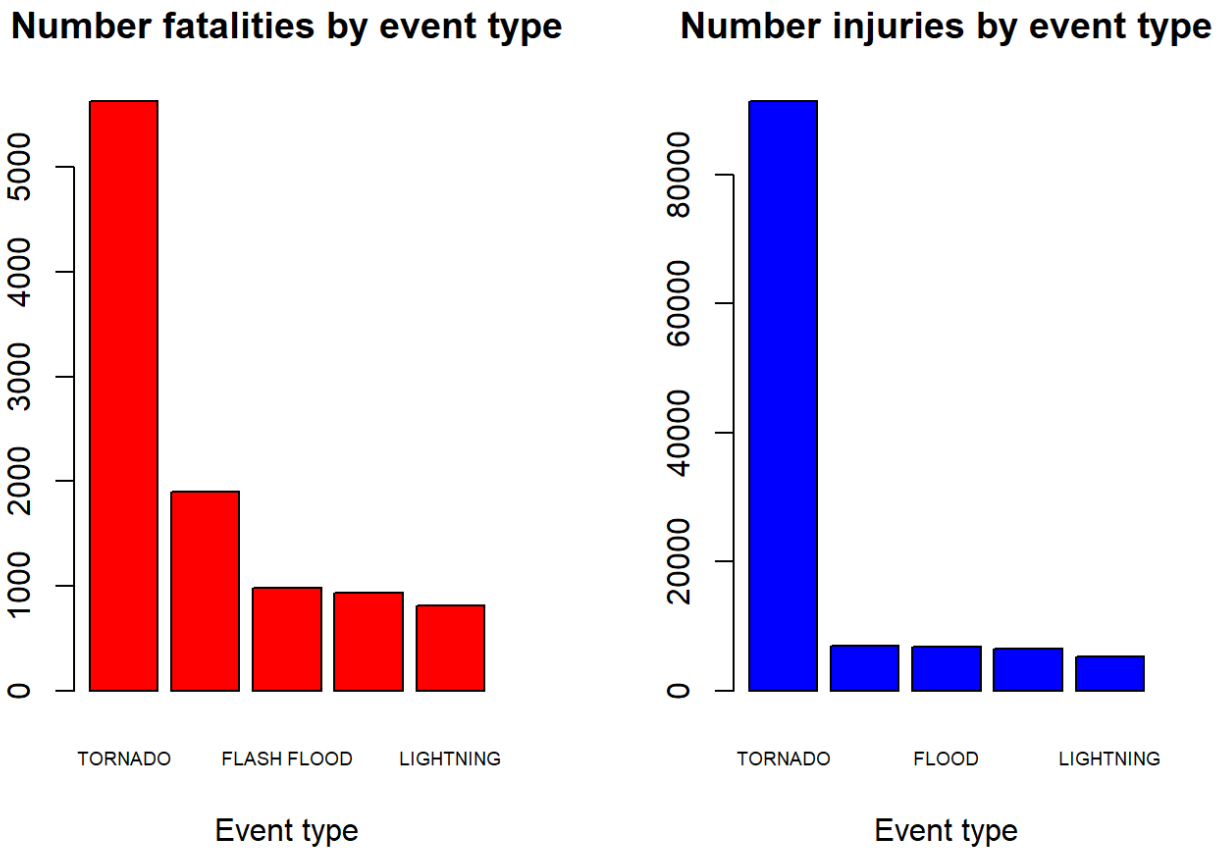
Results

In this chapter the results of our analysis are presented.

The first question we want answered is which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health?

```
par(mfrow=c(1,2))
barplot(fatalities$FATALITIES[1:5], main="Number fatalities by event type",
        xlab="Event type",
        horiz=FALSE,
        names.arg=fatalities$EVTYPE[1:5],
        cex.names=0.6,
        col="red")

barplot(injuries$INJURIES[1:5], main="Number injuries by event type",
        xlab="Event type",
        horiz=FALSE,
        names.arg=injuries$EVTYPE[1:5],
        cex.names=0.6,
        col="blue")
```



From the two plots above we can conclude that both in terms of fatalities as well as injuries, tornado's are by far the most harmfull to the populations health. Just for the fun of it, let's see which extent

```
round(injuries$INJURIES[1]/sum(injuries$INJURIES)*100,0)
```

```
## [1] 65
```

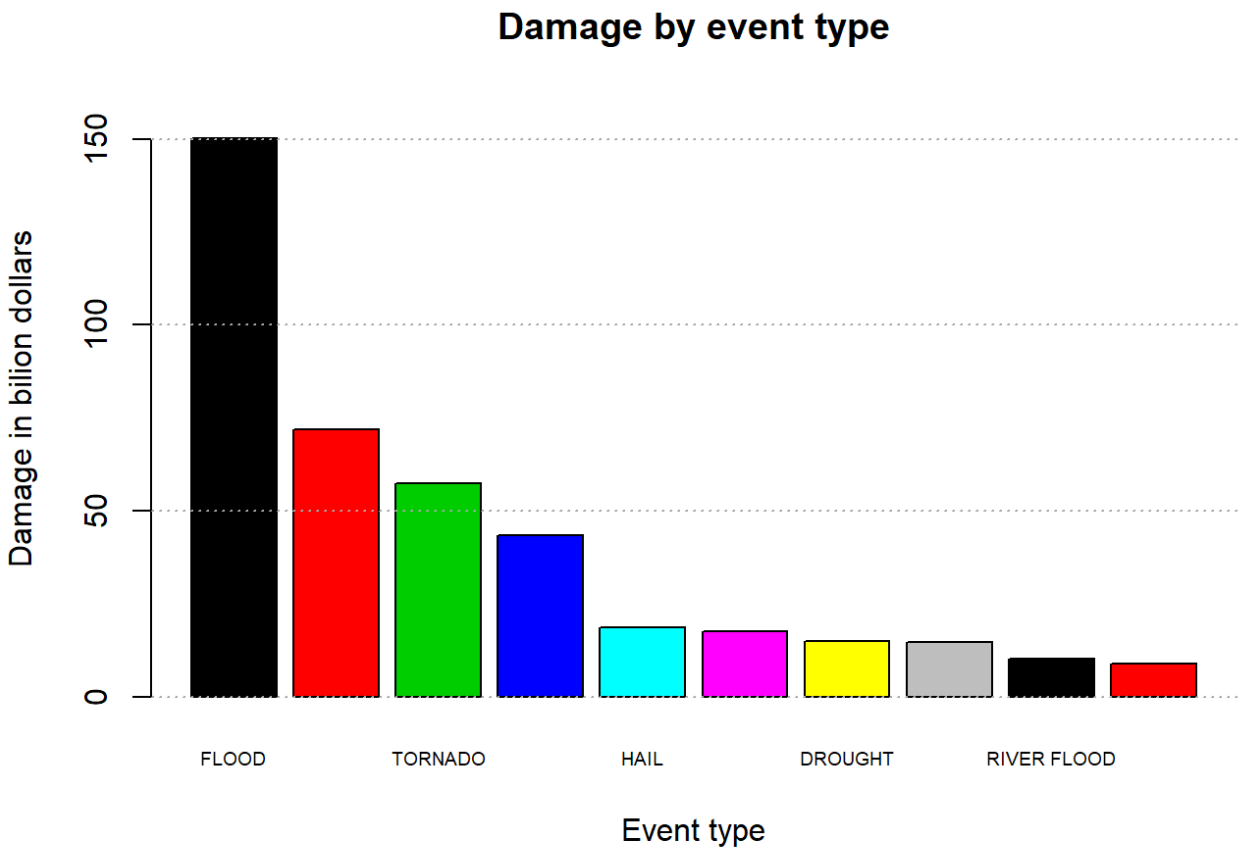
```
round(fatalities$FATALITIES[1]/sum(fatalities$FATALITIES)*100,0)
```

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

It turns out that tornado's are responsible for 65% injuries and 37% of fatalities caused by storms in North-America.

The second question was which types of events have the greatest economic consequences?

```
barplot(damage$`crop + prop`[1:10]/1e9, main="Damage by event type",
        xlab="Event type",
        ylab="Damage in bilion dollars",
        horiz=FALSE,
        names.arg=damage$EVTYPE[1:10],
        ylim=c(0,160),
        cex.names=0.6,
        col=1:10)
grid(nx = NA, ny = NULL, col = "darkgray", lty = "dotted",
     lwd = par("lwd"), equilogs = TRUE)
```




```
round(damage$`crop + prop`[1]/sum(damage$`crop + prop`)*100,0)
```

```
## [1] 32
```

Floods are responsible for 32% of total damage to property caused by storms in North-America.