Analysis of Storm dataset

Stefano Masneri

15 March 2016

Synopsis

The goal of this document is to analyze the NOAA Storm Dataset, containing data about weather in the US from the year 1950, to answer some questions about the effects of severe weather events. In particular the document will try to address the following questions:

1. Across the United States, which types of events are most harmful with respect to population health?
2. Across the United States, which types of events have the greatest economic consequences?

The *Data Processing* section shows how meaningful information is extracted from the raw data, while the *Results* section provides the answers to the previous questions.

Data Processing

Before we start

We load all the library we need for the analysis and set some global options

```
library(ggplot2)
```

Getting the data

We then proceed downloading the raw data and load it in our environment

```
if ( !file.exists('StormData.csv.bz2') ) {
  download.file('https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2',
                'StormData.csv.bz2'  )
}
storm <- read.csv( bzfile( 'StormData.csv.bz2' ) )
```

Before cleaning the data or running any analysis, we just peek into the content of the data

```
colnames(storm)
```

```
##  [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"
```

```
dim(storm)
```

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

Analyzing data

To answer the first question we have to aggregate the data with respect with the event. We decided to consider both fatalities and injuries in this case.

```
deainj <- aggregate(cbind(INJURIES, FATALITIES) ~ EVTYPE, storm, FUN = sum, na.rm = TRUE)
deainj$INJURIES_AND_DEATHS <- deainj$INJURIES + deainj$FATALITIES
head(deainj)
```

```
##                 EVTYPE INJURIES FATALITIES INJURIES_AND_DEATHS
## 1    HIGH SURF ADVISORY        0          0                   0
## 2         COASTAL FLOOD        0          0                   0
## 3           FLASH FLOOD        0          0                   0
## 4             LIGHTNING        0          0                   0
## 5             TSTM WIND        0          0                   0
## 6       TSTM WIND (G45)        0          0                   0
```

Preparing the damage information

There is some more work involved in order to extract useful information about the damages done by weather events. We have to consider both property and crop damages, and take into account that there is one column describing the "base" damage and another one describing the "order of magnitude" of the damage. Furthermore, some cleaning is needed, since some values just don't make any sense.

```
storm$PROPDMGEXP_CLEAN <- 0
storm$PROPDMGEXP_CLEAN[grep("k", storm$PROPDMGEXP,ignore.case = T)] <- 3
storm$PROPDMGEXP_CLEAN[grep("m", storm$PROPDMGEXP,ignore.case = T)] <- 6
storm$PROPDMGEXP_CLEAN[grep("b", storm$PROPDMGEXP,ignore.case = T)] <- 9
storm$PROPDMG_TOT <- 10^as.numeric(storm$PROPDMGEXP_CLEAN) * storm$PROPDMG

storm$CROPDMGEXP_CLEAN <- 0
storm$CROPDMGEXP_CLEAN[grep("k", storm$CROPDMGEXP,ignore.case = T)] <- 3
storm$CROPDMGEXP_CLEAN[grep("m", storm$CROPDMGEXP,ignore.case = T)] <- 6
storm$CROPDMGEXP_CLEAN[grep("b", storm$CROPDMGEXP,ignore.case = T)] <- 9
storm$CROPDMG_TOT <- 10^as.numeric(storm$CROPDMGEXP_CLEAN) * storm$CROPDMG
```

After cleaning the data we can aggregate it similarly to what we did for injuries and fatalities.

```
damage <- aggregate(cbind(CROPDMG_TOT, PROPDMG_TOT) ~ EVTYPE, storm, FUN = sum, na.rm = TRUE)
damage$CROP_AND_PROP <- damage$CROPDMG_TOT + damage$PROPDMG_TOT
head(damage)
```

```
##                 EVTYPE CROPDMG_TOT PROPDMG_TOT CROP_AND_PROP
## 1    HIGH SURF ADVISORY           0      200000        200000
## 2         COASTAL FLOOD           0           0             0
## 3           FLASH FLOOD           0       50000         50000
```

```
## 4             LIGHTNING           0          0            0
## 5             TSTM WIND           0    8100000      8100000
## 6       TSTM WIND (G45)           0       8000         8000
```

Results

Regarding the first question, we want to see which weather event caused the maximum amount of deaths and injuries

```
max_deaths <- max(deainj$FATALITIES)
max_injuries <- max(deainj$INJURIES)
max_combined <- max(deainj$INJURIES_AND_DEATHS)
event_max_deaths <- as.character(deainj[[which.max(deainj$FATALITIES), 'EVTYPE']])
event_max_injuries <- as.character(deainj[[which.max(deainj$INJURIES), 'EVTYPE']])
```

So we can see that the event which caused the most deaths is TORNADO and likewise for injuries the event is the same: TORNADO.

Regarding the second question, we proceed in a similar fashion:

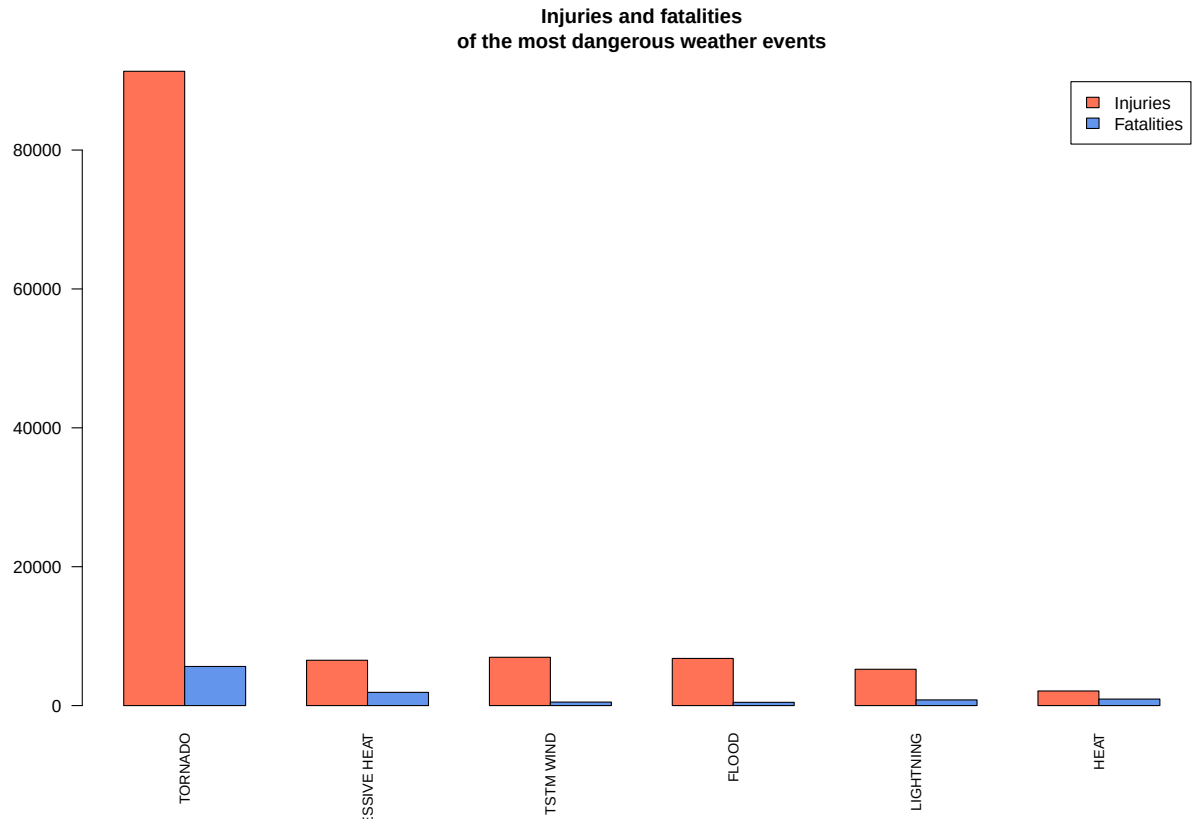
```
max_crop_dmg <- max(damage$CROPDMG_TOT)
max_prop_dmg <- max(damage$PROPDMG_TOT)
max_damage <- max(damage$CROP_AND_PROP)
event_max_crop_dmg <- as.character(damage[[which.max(damage$CROPDMG_TOT), 'EVTYPE']])
event_max_prop_dmg <- as.character(deainj[[which.max(damage$PROPDMG_TOT), 'EVTYPE']])
```

This shows that the event which caused the most damage to properties is FLOOD (for a total of 144.6577098 billion dollars), while for crop the most damage has been caused by DROUGHT (13.972566 billion dollars).

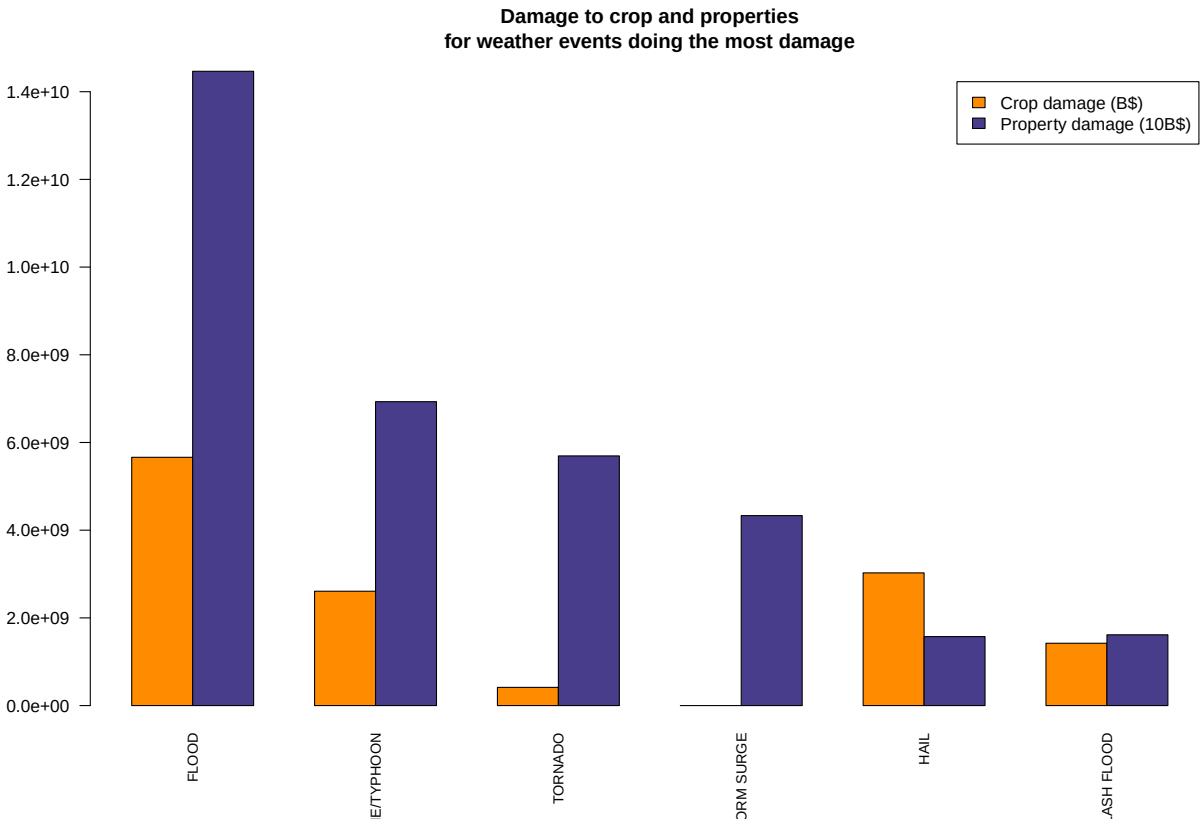
Finally, we show two plots describing the total amount of damage and the death + injuries caused by a selection of weather events. We just picked the events that caused the most damage or injuries. To do that we just sort the data according to the total.

```
sorted_deainj <- deainj[order(deainj$INJURIES_AND_DEATHS, decreasing = TRUE), ]
sorted_damage <- damage[order(damage$CROP_AND_PROP, decreasing = TRUE), ]
plot_deainj <- matrix(c(head(sorted_deainj$INJURIES), head(sorted_deainj$FATALITIES)),
                      nrow=2, byrow=T)
plot_damage <- matrix(c(head(sorted_damage$CROPDMG_TOT), head(sorted_damage$PROPDMG_TOT/10)),
                      nrow=2, byrow=T)

barplot( plot_deainj, beside = TRUE, col=c("coral1","cornflowerblue"),
         cex.names = 0.8, las = 2, legend=c("Injuries","Fatalities"),
         names.arg = head(sorted_deainj$EVTYPE),
         main = "Injuries and fatalities\nof the most dangerous weather events")
```




```
barplot( plot_damage, beside = TRUE, col=c("darkorange","darkslateblue"),
         cex.names = 0.8, las = 2, legend=c("Crop damage (B$)","Property damage (10B$)"),
         names.arg = head(sorted_damage$EVTYPE),
         main = "Damage to crop and properties\nfor weather events doing the most damage")
```



We can derive a few interesting things from the plots. For example we notice that damages to crops and properties look quite unrelated, and that the damages to properties are far bigger than the ones to crops. For the first plot, we see that the number of injuries is always bigger than the number of fatalities (as intuitively makes sense). Finally, we see that tornado and flood appear in both flood, meaning that they are events which cause both economic and health damage.