StormData.Rmd

Polina Filipova

May 14, 2017

Overview

This document is generated for an asignment under the Reproducible Research course, offered by Johns Hopkins University on Coursera.

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.

The following questions are addressed:

- **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?**

The data for this assignment come in the form of a comma-separated-value file compressed via the bzip2 algorithm to reduce its size. It can be downloaded from the course web site:

- Storm data

See also the NOAA documentation:

- Storm data FAQ page
- Storm data preparation

The following required items can be reviewed below:

1. Code for reading the dataset
2. Data transformation justifications & Code for processing the data
3. Results and conclusions
4. All of the R code needed to reproduce the results (numbers, plots, etc.)

1. Read and Review Data

Note: This will search for content in the current working directory for your R environment. R is capable of reading compressed .csv.

```
# Check if we already have the data. If not, fetch and extract it:

if(!file.exists("repdata%2Fdata%2FStormData.csv.bz2"))
        {
download.file("https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2", destfile = "r
        }
```

```
# read.csv is capable of reading .bz compression and we will make use of that here:
```

```r
stormData <- read.csv("repdata%2Fdata%2FStormData.bz2", header=TRUE, sep=",", stringsAsFactor=FALSE, na
str(stormData)
```

```
## 'data.frame':    902297 obs. of  37 variables:
##  $ STATE__   : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ BGN_DATE  : chr  "4/18/1950 0:00:00" "4/18/1950 0:00:00" "2/20/1951 0:00:00" "6/8/1951 0:00:00" .
##  $ BGN_TIME  : chr  "0130" "0145" "1600" "0900" ...
##  $ TIME_ZONE : chr  "CST" "CST" "CST" "CST" ...
##  $ COUNTY    : num  97 3 57 89 43 77 9 123 125 57 ...
##  $ COUNTYNAME: chr  "MOBILE" "BALDWIN" "FAYETTE" "MADISON" ...
##  $ STATE     : chr  "AL" "AL" "AL" "AL" ...
##  $ EVTYPE    : chr  "TORNADO" "TORNADO" "TORNADO" "TORNADO" ...
##  $ BGN_RANGE : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BGN_AZI   : chr  "" "" "" "" ...
##  $ BGN_LOCATI: chr  "" "" "" "" ...
##  $ END_DATE  : chr  "" "" "" "" ...
##  $ END_TIME  : chr  "" "" "" "" ...
##  $ COUNTY_END: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ COUNTYENDN: logi  NA NA NA NA NA NA ...
##  $ END_RANGE : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ END_AZI   : chr  "" "" "" "" ...
##  $ END_LOCATI: chr  "" "" "" "" ...
##  $ LENGTH    : num  14 2 0.1 0 0 1.5 1.5 0 3.3 2.3 ...
##  $ WIDTH     : num  100 150 123 100 150 177 33 33 100 100 ...
##  $ F         : int  3 2 2 2 2 2 2 1 3 3 ...
##  $ MAG       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ FATALITIES: num  0 0 0 0 0 0 0 0 1 0 ...
##  $ INJURIES  : num  15 0 2 2 2 6 1 0 14 0 ...
##  $ PROPDMG   : num  25 2.5 25 2.5 2.5 2.5 2.5 2.5 25 25 ...
##  $ PROPDMGEXP: chr  "K" "K" "K" "K" ...
##  $ CROPDMG   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CROPDMGEXP: chr  "" "" "" "" ...
##  $ WFO       : chr  "" "" "" "" ...
##  $ STATEOFFIC: chr  "" "" "" "" ...
##  $ ZONENAMES : chr  "" "" "" "" ...
##  $ LATITUDE  : num  3040 3042 3340 3458 3412 ...
##  $ LONGITUDE : num  8812 8755 8742 8626 8642 ...
##  $ LATITUDE_E: num  3051 0 0 0 0 ...
##  $ LONGITUDE_: num  8806 0 0 0 0 ...
##  $ REMARKS   : chr  "" "" "" "" ...
##  $ REFNUM    : num  1 2 3 4 5 6 7 8 9 10 ...
```

```r
dim(stormData)
```

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

For this study, our interest lies with event type, begin/end time, fatalities, injuries, crop and property damage. For clarity, we can convert BGN_DATE to POSIXlt format.

```r
stormData$BGN_DATE <- as.character(stormData$BGN_DATE)
stormData$BGN_DATE <- as.Date(stormData$BGN_DATE, "%m/%d/%Y %H:%M:%S")
head(stormData$BGN_DATE)
```

```
## [1] "1950-04-18" "1950-04-18" "1951-02-20" "1951-06-08" "1951-11-15"
## [6] "1951-11-15"
```

We will drop fields beyond the scope of this study, as we go along.

Finally, the libraries in use are:

```
if(!require(dplyr)) { install.packages("dplyr") }
```

```
## Loading required package: 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(dplyr)

if(!require(ggplot2)) { install.packages("ggplot2") }
```

```
## Loading required package: ggplot2
```

```
library(ggplot2)
```

Back to Overview

2. Data transformation justifications & Code for processing the data

Calculating Human Life Damages

Here we look at INJURIES and FATALITIES, and we will combine them in one column, HARM. The NOAA database includes storm data beginning in 1950, and it is fairly prone to lapses up until the mid-1990s, something to consider as well.

```
stormData1 <-
    stormData %>%
    group_by(EVTYPE) %>%
    select(EVTYPE, BGN_DATE, FATALITIES, INJURIES, PROPDMG, PROPDMGEXP, CROPDMG, CROPDMGEXP) %>%
    mutate(HARM = FATALITIES + INJURIES) %>%
    filter(BGN_DATE > "1994-12-31")

mean(is.na(stormData1))
```

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

Great - we have no N/A values here.

Calculating Economical Damages

From the FAQ, we see that the damages are a compound of PROPDMG by PROPDMGEXP for property, CROPDMG by CROPDMGEXP for crops. Above, we saw the exponential represented in "K". The remainder of the unique exponential types are:

```
uniqPEXP <- unique(stormData$PROPDMGEXP)
uniqPEXP
```

```
##  [1] "K" "M" ""  "B" "m" "+" "0" "5" "6" "?" "4" "2" "3" "h" "7" "H" "-"
## [18] "1" "8"
```

```r
uniCPEXP <- unique(stormData$CROPDMGEXP)
uniCPEXP
```

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

This confirms the NOAA standard of K/k for thousands, M/m for millions, b/B for billions and we see some stray values also. Let's ensure we work with valid values only, and in one and the same format. The rest of the exponentials would not be statistically significant, in comparison.

```r
# R does not give us an easy out when it comes to ignoring case,
# unless we go into regular expressions with grep(l).
# Fortunately, we have just a few values of this sort.

convertValues <- function(value, EXP)
    {
  if (EXP == "B" || EXP == "b")
      {
    new.value = value * 10**9
      }
  if (EXP == "M" || EXP == "m")
      {
    new.value =  value * 10**6
      }
  if (EXP == "K" | EXP == "k")
      {
    new.value = value * 10**3
      }
  new.value
}

convertValuesVect <- Vectorize(convertValues)

# Dropping the scientific notation for the values.

format(convertValuesVect, scientific = FALSE)

# We can group the damage cost per type in a single column.

validTypes <- c("B", "b", "M", "m", "K", "k")

stormData2 <- stormData1 %>%
  select(EVTYPE, BGN_DATE, HARM, PROPDMG, PROPDMGEXP, CROPDMG, CROPDMGEXP) %>%
  filter(PROPDMGEXP %in% validTypes & CROPDMGEXP %in% validTypes) %>%
  mutate(CROPCOST = convertValuesVect(CROPDMG, CROPDMGEXP)) %>%
  mutate(PROPCOST = convertValuesVect(PROPDMG, PROPDMGEXP)) %>%
  mutate(DMGCOST = CROPCOST + PROPCOST)

stormDataFin <- stormData2 %>%
  select(EVTYPE, BGN_DATE, HARM, DMGCOST)
```

Back to Overview

3. Results and conclusions

Displaying Human Life Damages and Economical Damages

```r
stormDataEV <- group_by(stormDataFin, EVTYPE)
stormDataEVtotal <- data.frame(summarise(stormDataEV, totalHARM = sum(HARM), totalCOST = sum(DMGCOST)))
top20stormDataHARM <- head(arrange(stormDataEVtotal, desc(totalHARM)), 20)
top20stormDataCOST <- head(arrange(stormDataEVtotal, desc(totalCOST)), 20)
```

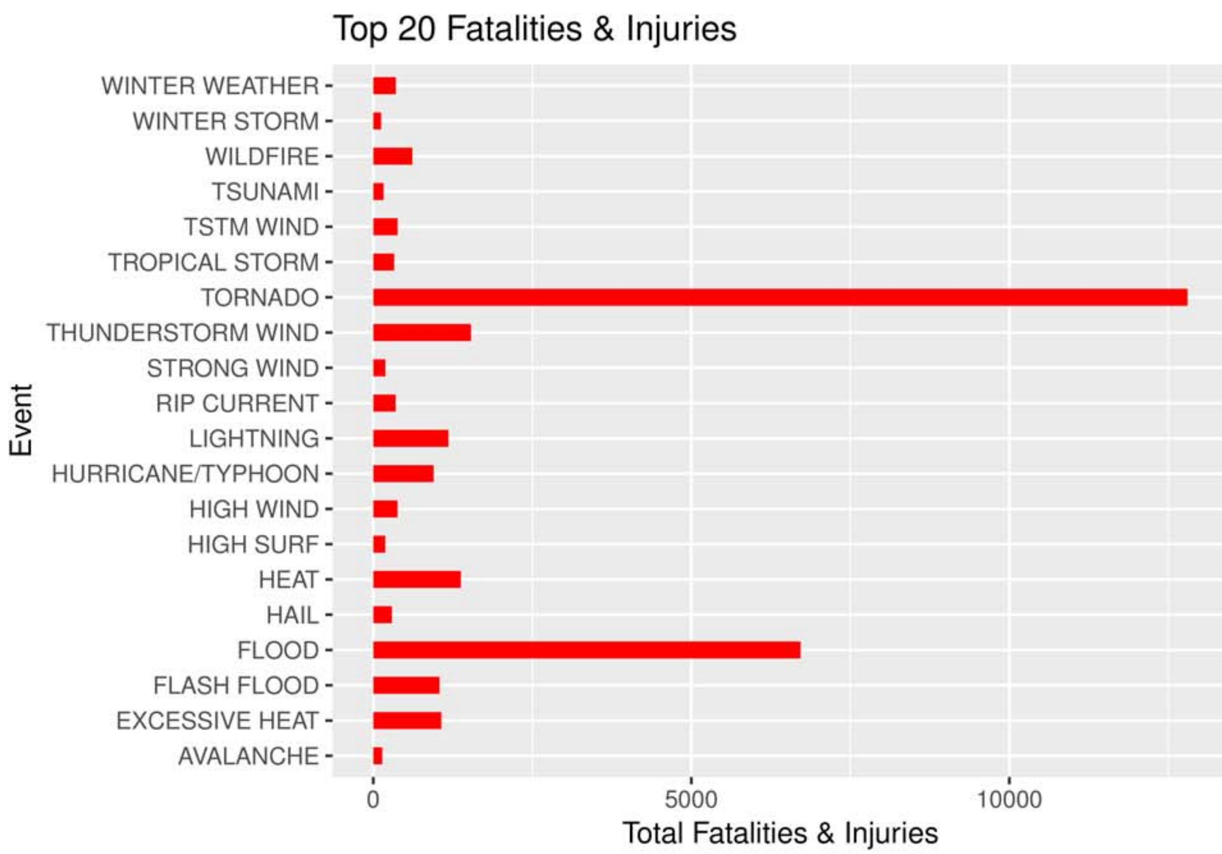
Damage to Human Life

Tornadoes are responsible for the most human life casualties on US soil since 1995.

```r
top20stormDataHARM
```

```
##                EVTYPE totalHARM    totalCOST
## 1             TORNADO     12803  16347960400
## 2               FLOOD      6718 137439035900
## 3   THUNDERSTORM WIND      1536   3811985440
## 4                HEAT      1376      2390000
## 5           LIGHTNING      1182    317965530
## 6      EXCESSIVE HEAT      1070    493803200
## 7         FLASH FLOOD      1041   8449990030
## 8   HURRICANE/TYPHOON       949  29348167800
## 9            WILDFIRE       614   3684468370
## 10          TSTM WIND       383   1155040110
## 11          HIGH WIND       382   3057106640
## 12     WINTER WEATHER       354     34897500
## 13        RIP CURRENT       351         1000
## 14     TROPICAL STORM       328   1507237350
## 15               HAIL       292   9519840090
## 16        STRONG WIND       190    184200560
## 17          HIGH SURF       188     83017500
## 18            TSUNAMI       162    144082000
## 19          AVALANCHE       141      2385800
## 20       WINTER STORM       123   1016068200
```

```r
gH <- ggplot(data = top20stormDataHARM, aes(x = totalHARM, y = EVTYPE))
gH + geom_segment(aes(xend = 0, yend = EVTYPE), size = 3, color = "red") + labs(x = "Total Fatalities &
```



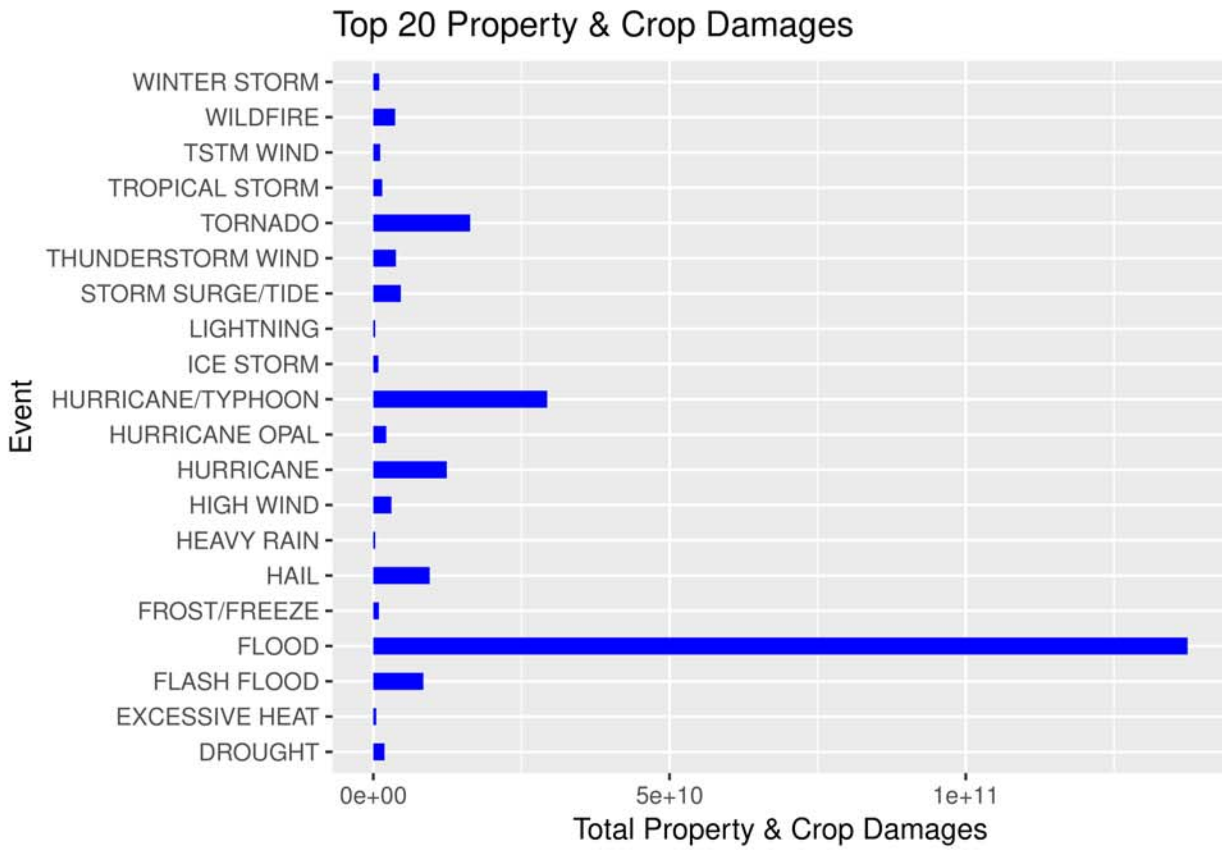
Damage to the Economy

Floods lead the crop and property damage costs in the USA since 1995.

```
top20stormDataCOST
```

```
##               EVTYPE totalHARM    totalCOST
## 1              FLOOD      6718 137439035900
## 2  HURRICANE/TYPHOON       949  29348167800
## 3            TORNADO     12803  16347960400
## 4          HURRICANE        64  12404268000
## 5               HAIL       292   9519840090
## 6        FLASH FLOOD      1041   8449990030
## 7   STORM SURGE/TIDE        16   4641493000
## 8  THUNDERSTORM WIND      1536   3811985440
## 9           WILDFIRE       614   3684468370
## 10         HIGH WIND       382   3057106640
## 11    HURRICANE OPAL         1   2187000000
## 12           DROUGHT         4   1886417000
## 13    TROPICAL STORM       328   1507237350
## 14         TSTM WIND       383   1155040110
## 15      WINTER STORM       123   1016068200
## 16      FROST/FREEZE         0    941281000
## 17         ICE STORM        28    862652300
## 18    EXCESSIVE HEAT      1070    493803200
## 19        HEAVY RAIN        61    320287730
## 20         LIGHTNING      1182    317965530
```

```
gC <- ggplot(data = top20stormDataCOST, aes(x = totalCOST, y = EVTYPE))
gC + geom_segment(aes(xend = 0, yend = EVTYPE), size = 3, color = "blue") + labs(x = "Total Property & C
```



Back to Overview

5. All of the R code needed to reproduce the results (numbers, plots, etc.)

Please refer to this GitHub location:

- https://github.com/VoidHamlet/NOAAStormData

Back to Overview