Reproducible Research Course Project 2

Ignacio Ojea

July 26, 2016

Report on the Population Health and Economic Impact Of Severe Weather Events in US

Introduction

This project is based on information provided by the National Oceanic and Atmospheric Administration's (NOAA) storm database. This encompasses 37 variables regarding the major weather events in the United States, including type of event, location, time, and in particular an estimation of harm to the population (fatalities and injuries) and to the economy (property and crop damage). The events in the database start in the year 1950 and end in November 2011. The purpose of this analysis is to identify the weather events that inflicted the most human and economic damage.

Synopsis

The analysis of the database revealed that tornados were the most damaging weather events to the population's health - for a total of 97043 between injuries and fatalities. The second worse event in this respect was excessive heat - but quite far away, for a total of 12421 between injuries and fatalities.

With respect to economic consequences, floods were the most significant for a total of $180463144933 in damages between property and crop damages. On second place there are the hurricanes, for a total damage of $90251472810.

The analysis proceeded by (i) downloading the data, (ii) subsetting it and cleaning it for the relevant variables, (iii) aggregating the data so that it is easy to process and plotting it.

Data Processing

Part 1: Loading and preprocessing the data

The first step consists in (a) downloading the file, (b) unzipping it, and (c) read the .cvs into the data table.

```r
download.file("https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2", dest="temp.bz
data <- read.csv(bzfile("temp.bz2"), header=TRUE, sep=",", stringsAsFactors=FALSE)
```

Part 2: Cleaning and subsetting the data

Basics

To simplyfy, I subset only the relevant variables with value > 0: the event types (variable “EVTYPE”), the figures related to population health impacts (variables “Fatalities” and “Injures”“), and the ones corresponding to the economic consequences (variables”PropDMG“,”PROPDMGEXP“,”CROPDMG" & “CROPDMGEXP”):

```r
data2 <- data[,c("EVTYPE", "FATALITIES", "INJURIES", "PROPDMG", "PROPDMGEXP", "CROPDMG", "CROPDMGEXP")]
```

Take a quick look at the data available.

```r
summary(data2)
```

```
##     EVTYPE            FATALITIES          INJURIES
##  Length:902297      Min.   :  0.0000   Min.   :   0.0000
##  Class :character   1st Qu.:  0.0000   1st Qu.:   0.0000
##  Mode  :character   Median :  0.0000   Median :   0.0000
##                     Mean   :  0.0168   Mean   :   0.1557
##                     3rd Qu.:  0.0000   3rd Qu.:   0.0000
##                     Max.   :583.0000   Max.   :1700.0000
##     PROPDMG         PROPDMGEXP            CROPDMG          CROPDMGEXP
##  Min.   :   0.00   Length:902297      Min.   :  0.000   Length:902297
##  1st Qu.:   0.00   Class :character   1st Qu.:  0.000   Class :character
##  Median :   0.00   Mode  :character   Median :  0.000   Mode  :character
##  Mean   :  12.06                      Mean   :  1.527
##  3rd Qu.:   0.50                      3rd Qu.:  0.000
##  Max.   :5000.00                      Max.   :990.000
```

Before anything else, notice that the event type variable and some of the economic variables need cleaning and formatting:

```r
length(unique(data2$EVTYPE))
```

```
## [1] 985
```

```r
unique(data2$PROPDMGEXP)
```

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

```r
unique(data2$CROPDMGEXP)
```

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

Cleaning the data

Let us start with the economic variables:

```
data2$PROP.DMG.EXP[(data2$PROPDMGEXP=='')|(data2$PROPDMGEXP=='-')|(data2$PROPDMGEXP=='?')|(data2$PROPDM
data2$PROP.DMG.EXP[(data2$PROPDMGEXP=='1')] <- 1
data2$PROP.DMG.EXP[(data2$PROPDMGEXP=='2')] <- 2
data2$PROP.DMG.EXP[(data2$PROPDMGEXP=='3')] <- 3
data2$PROP.DMG.EXP[(data2$PROPDMGEXP=='4')] <- 4
data2$PROP.DMG.EXP[(data2$PROPDMGEXP=='5')] <- 5
data2$PROP.DMG.EXP[(data2$PROPDMGEXP=='6')] <- 6
data2$PROP.DMG.EXP[(data2$PROPDMGEXP=='7')] <- 7
data2$PROP.DMG.EXP[(data2$PROPDMGEXP=='8')] <- 8
data2$PROP.DMG.EXP[(data2$PROPDMGEXP=='H') | (data2$PROPDMGEXP=='h')] <- 2
data2$PROP.DMG.EXP[(data2$PROPDMGEXP=='K') | (data2$PROPDMGEXP=='k')] <- 3
data2$PROP.DMG.EXP[(data2$PROPDMGEXP=='M')] <- 6
data2$PROP.DMG.EXP[(data2$PROPDMGEXP=='B')] <- 9

data2$CROP.DMG.EXP[(data2$CROPDMGEXP=='')|(data2$CROPDMGEXP=='-')|(data2$CROPDMGEXP=='?')|(data2$CROPDM
data2$CROP.DMG.EXP[(data2$CROPDMGEXP=='K')|(data2$CROPDMGEXP=='k')] <- 3
data2$CROP.DMG.EXP[(data2$CROPDMGEXP=='M')|(data2$CROPDMGEXP=='m')] <- 6
data2$CROP.DMG.EXP[(data2$CROPDMGEXP=='B')] <- 9

#Now find the total cost of property damage
data2$PROP.DMG.COST <- data2$PROPDMG*10^as.numeric(data2$PROP.DMG.EXP)
data2$CROP.DMG.COST <- data2$CROPDMG*10^as.numeric(data2$CROP.DMG.EXP)
```

Now allow me to clean up a little bit the variable on event types:

```
# I start by setting everything to upper case letters
data2$EVTYPE <- toupper(data2$EVTYPE)

# Then I group categories according to their name
data2$EVTYPE <- gsub('.*LOW.*TEMPER.*', 'COLD', data2$EVTYPE)
data2$EVTYPE <- gsub('.*HIGH.*TEMPER.*', 'HEAT', data2$EVTYPE)
data2$EVTYPE <- gsub('.*HEAT.*', 'HEAT', data2$EVTYPE)
data2$EVTYPE <- gsub('.*WARM.*', 'HEAT', data2$EVTYPE)
data2$EVTYPE <- gsub('.*HIGH.*TEMP.*', 'EXTREME HEAT', data2$EVTYPE)
data2$EVTYPE <- gsub('.*.*RECORD HIGH TEMPERATURES.*', 'EXTREME HEAT', data2$EVTYPE)
data2$EVTYPE <- gsub('.*FIRE.*', 'FIRE', data2$EVTYPE)
data2$EVTYPE <- gsub('.*HURRICANE.*', 'HURRICANE', data2$EVTYPE)
data2$EVTYPE <- gsub('.*RAIN.*', 'RAIN', data2$EVTYPE)
data2$EVTYPE <- gsub('.*STORM.*', 'STORM', data2$EVTYPE)
data2$EVTYPE <- gsub('.*FLOOD.*', 'FLOOD', data2$EVTYPE)
data2$EVTYPE <- gsub('.*WIND.*', 'WIND', data2$EVTYPE)
data2$EVTYPE <- gsub('.*WND.*', 'WIND', data2$EVTYPE)
data2$EVTYPE <- gsub('.*TORN.*', 'TORNADO', data2$EVTYPE)
data2$EVTYPE <- gsub('.*HAIL.*', 'HAIL', data2$EVTYPE)
data2$EVTYPE <- gsub('.*SNOW.*', 'SNOW', data2$EVTYPE)
data2$EVTYPE <- gsub('.*CLOUD.*', 'CLOUD', data2$EVTYPE)
data2$EVTYPE <- gsub('.*MICROBURST.*', 'MICROBURST', data2$EVTYPE)
data2$EVTYPE <- gsub('.*BLIZZARD.*', 'BLIZZARD', data2$EVTYPE)
data2$EVTYPE <- gsub('.*COLD.*', 'COLD', data2$EVTYPE)
data2$EVTYPE <- gsub('.*SNOW.*', 'COLD', data2$EVTYPE)
```

```
data2$EVTYPE <- gsub('.*FREEZ.*', 'COLD', data2$EVTYPE)
data2$EVTYPE <- gsub('.*LOW TEMPERATURE RECORD.*', 'COLD', data2$EVTYPE)
data2$EVTYPE <- gsub('.*ICE.*', 'COLD', data2$EVTYPE)
data2$EVTYPE <- gsub('.*FROST.*', 'COLD', data2$EVTYPE)
data2$EVTYPE <- gsub('.*LO.*TEMP.*', 'COLD', data2$EVTYPE)
data2$EVTYPE <- gsub('.*FROST.*', 'COLD', data2$EVTYPE)
data2$EVTYPE <- gsub('.*HIGH.*TEMPER.*', 'HEAT', data2$EVTYPE)
data2$EVTYPE <- gsub('.*TORNADO.*', 'TORNADO', data2$EVTYPE)
data2$EVTYPE <- gsub('.*DRY.*', 'DRY', data2$EVTYPE)
data2$EVTYPE <- gsub('.*DUST.*', 'DUST', data2$EVTYPE)
data2$EVTYPE <- gsub('.*RAIN.*', 'RAIN', data2$EVTYPE)
data2$EVTYPE <- gsub('.*LIGHTNING.*', 'LIGHTNING', data2$EVTYPE)
data2$EVTYPE <- gsub('.*SUMMARY.*', 'SUMMARY', data2$EVTYPE)
data2$EVTYPE <- gsub('.*WET.*', 'WET', data2$EVTYPE)
data2$EVTYPE <- gsub('.*FIRE.*', 'FIRE', data2$EVTYPE)
data2$EVTYPE <- gsub('.*FOG.*', 'FOG', data2$EVTYPE)
data2$EVTYPE <- gsub('.*VOLCANIC.*', 'VOLCANIC', data2$EVTYPE)
data2$EVTYPE <- gsub('.*SURF.*', 'SURF', data2$EVTYPE)
```

Lets take a look at the new amount of event types:

```
length(unique(data2$EVTYPE))
```

```
## [1] 147
```

Looking good!

Part 3: Analyzing (aggregating and plotting) the data

Reshaping the data

Now lets generate two data sets, each with the relevant information to answer one of the questions (considering only the cases in which at least one of the data is different from zero):

```
#Let me start with health.
Fatalities.data <- aggregate(x = list(FATALITIES = data2$FATALITIES), by=list(EVENT.TYPE=data2$EVTYPE),

Injuries.data <- aggregate(x = list(INJURIES = data2$INJURIES), by=list(EVENT.TYPE=data2$EVTYPE), FUN=su

Health.data <- merge(Fatalities.data,Injuries.data, by="EVENT.TYPE")

Health.data$TOTAL.DAMAGE <- (Health.data$FATALITIES + Health.data$INJURIES)

Health.data <- Health.data[, c("EVENT.TYPE","TOTAL.DAMAGE", "INJURIES","FATALITIES")]

Health.data <- Health.data[order(Health.data$TOTAL.DAMAGE, decreasing=T),]

#Now for the economic damage.
```

```r
Property.damage.data <- aggregate(x = list(PROP.DMG = data2$PROP.DMG.COST), by=list(EVENT.TYPE=data2$EV

Crop.damage.data <- aggregate(x = list(CROP.DMG = data2$CROP.DMG.COST), by=list(EVENT.TYPE=data2$EVTYPE)

Economic.data <- merge(Property.damage.data, Crop.damage.data, by = "EVENT.TYPE")

Economic.data$TOTAL.DMG <- (Economic.data$PROP.DMG + Economic.data$CROP.DMG)

Economic.data <- Economic.data[, c("EVENT.TYPE", "TOTAL.DMG", "PROP.DMG", "CROP.DMG")]

Economic.data <- Economic.data[order(Economic.data$TOTAL.DMG, decreasing=T),]
```

Plotting the data

Basics

Now for plotting the results. I need some libraries for this part:

```r
library(ggplot2)
library(reshape2)
```

Let us now look at the top ten events in each category, ordered by the total damage they inflicted.

```r
head(Health.data,10)
```

```
##     EVENT.TYPE TOTAL.DAMAGE INJURIES FATALITIES
## 115    TORNADO        97043    91407       5636
## 39        HEAT        12421     9243       3178
## 141       WIND        10281     9044       1237
## 31       FLOOD        10126     8602       1524
## 111      STORM         7325     6692        633
## 67   LIGHTNING         6048     5231        817
## 16        COLD         1975     1581        394
## 29        FIRE         1698     1608         90
## 57   HURRICANE         1463     1328        135
## 38        HAIL         1386     1371         15
```

```r
head(Economic.data,10)
```

```
##     EVENT.TYPE    TOTAL.DMG     PROP.DMG    CROP.DMG
## 31       FLOOD 180463144933 168196218833 12266926100
## 57   HURRICANE  90251472810  84736180010  5515292800
## 111      STORM  79668064754  73261145866  6406918888
## 115    TORNADO  57406779946  56991818426   414961520
## 38        HAIL  18777980986  15731143513  3046837473
## 23     DROUGHT  15018672000   1046106000 13972566000
## 141       WIND  13740435768  12344216618  1396219150
## 29        FIRE   8904910130   8501628500   403281630
## 16        COLD   4714975050   1174134650  3540840400
## 88        RAIN   4189545992   3270230192   919315800
```

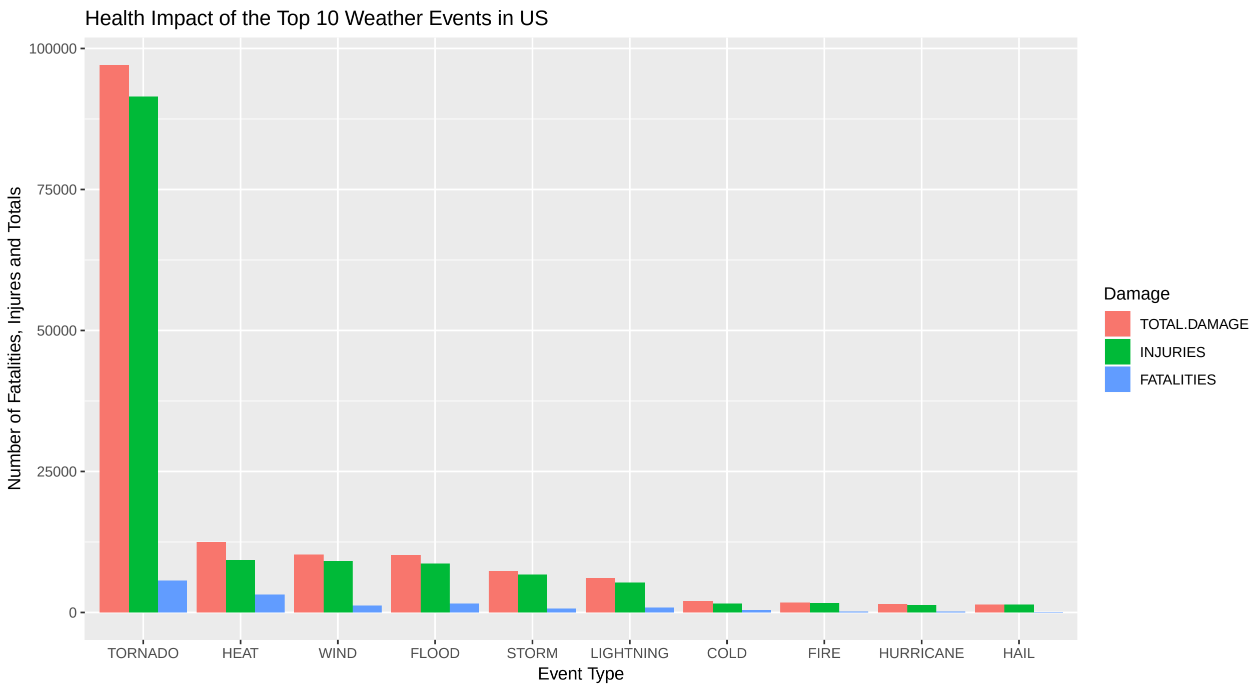
Plotting

Now for the plotting. First the Health Chart:

```
Temp.health.data <- melt(head(Health.data, 10), id.vars="EVENT.TYPE")

Health.Chart <- ggplot(Temp.health.data, aes(x=reorder(EVENT.TYPE, -value), y=value, fill = Damage)) +

print(Health.Chart)
```

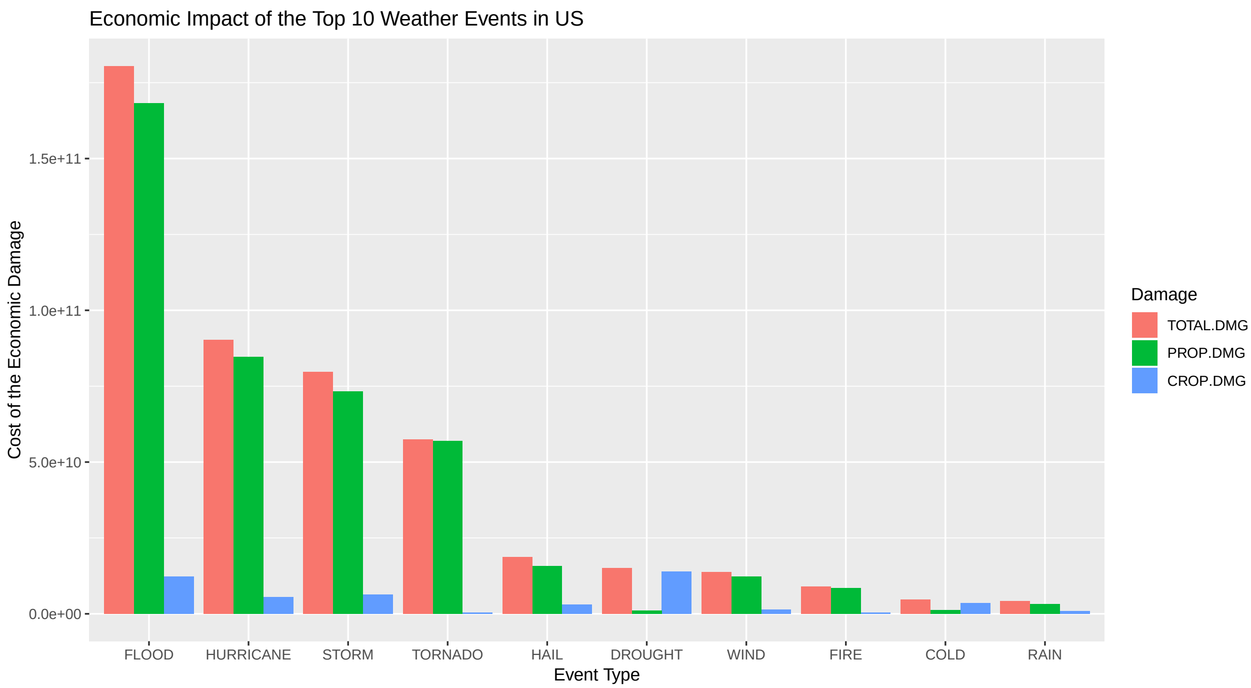


And now for the Economic Chart:

```
Temp.econ.data <- melt(head(Economic.data, 10), id.vars="EVENT.TYPE")

Economic.Chart <- ggplot(Temp.econ.data, aes(x=reorder(EVENT.TYPE, -value), y=value, fill = Damage)) +

print(Economic.Chart)
```



Results

The analysis of the database revealed that **tornados** were the most damaging weather events to the population's health - for a total of 97043 between injuries and fatalities. The second worse event in this respect was excessive **heat** - but quite far away, for a total of 12421 between injuries and fatalities. Excessive Wind (10281), floods (10126) and storms (7325) follow respectively.

With respect to economic impact, **floods** were the most significant for a total of $180463144933 in damages between property and crop damages. On second place there are the **hurricanes**, for a total damage of $90251472810. Storms ($79668064754), tornados ($57406779946) and hail ($18777980986) follow respectively.