Weather events wich resulted in most population health and econimic damage

Alexandr Cherkashin

Wednesday, January 23, 2015

## Synopsis

This project will address some basic questions about severe weather events:

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

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 events in the database start in the year 1950 and end in November 2011.

We will find that...

## Data Processing

### Used r packages

For further analisys and data transformations will use several r packages. Check that needed packages had been installed and install them if not:

list.of.packages <- c("ggplot2", "dplyr", "tidyr", "scales")  
new.packages <- list.of.packages[!(list.of.packages %in% installed.packages()[,"Package"])]  
if(length(new.packages)) install.packages(new.packages)  
sapply(list.of.packages, require, quietly = TRUE, character.only = TRUE)

## ggplot2 dplyr tidyr scales   
## TRUE TRUE TRUE TRUE

#remove unused variables  
rm(list.of.packages)  
rm(new.packages)

### Data source

The data for this project come in the form of a comma-separated-value file compressed via the bzip2 algorithm to reduce its size. It comes from course web site by this link: <http://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2>

Download data unless it's alredy done:

if (!file.exists("./data")) {dir.create("./data")}  
if (!file.exists("./data/StormData.bz2")) {download.file("http://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2", "./data/StormData.bz2", mode = "wb")}

### Data preparation

Extract and read the data

con <- bzfile("./data/StormData.bz2", open="r")  
data <- read.csv(con)  
close(con)  
rm(con)

Look at colum names:

colnames(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 the purpose of this project we need only some, so get rid of others to make dataset smaller.

data <- select(data, EVTYPE, 23:28)

According to [Storm Data Documentation](https://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2Fpd01016005curr.pdf) section 2.7:

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. <...> If the dollar amount of damage is unknown, or not available, check the “no information available” box.

Lets look at PROPDMGEXP and CROPDMGEXP colums that contains these magnitude signifyers:

table(data[data$PROPDMG!=0, "PROPDMGEXP"])

##   
## - ? + 0 1 2 3 4 5   
## 76 1 0 5 209 0 1 1 4 18   
## 6 7 8 B h H K m M   
## 3 2 0 40 1 6 227481 7 11319

table(data[data$CROPDMG!=0, "CROPDMGEXP"])

##   
## ? 0 2 B k K m M   
## 3 0 12 0 7 21 20137 1 1918

For this analisys we'll assume that lower case characters means that capital ones do. As we can see, there are several signifyers of unkown type like "h" or "?" but they have very small accurance when the corresponding DMG colums is not equal zero. There could be different theories but we will not do any bold assumtions. Accrording to such a small occurance We will map such values to zero and then multiply damage columns by it. We will use summary (not mean) aggregation later, so it is ok for now.

#make all chars same size  
data$PROPDMGEXP <- tolower(as.character(data$PROPDMGEXP))  
data$CROPDMGEXP <- tolower(as.character(data$CROPDMGEXP))  
  
#making multiplyer  
data$PROPDMGEXP[data$PROPDMGEXP!="b" & data$PROPDMGEXP!="m" & data$PROPDMGEXP!="k"] <- 0  
data$PROPDMGEXP[data$PROPDMGEXP=="b"] <- 1000000000  
data$PROPDMGEXP[data$PROPDMGEXP=="m"] <- 1000000  
data$PROPDMGEXP[data$PROPDMGEXP=="k"] <- 1000  
  
data$CROPDMGEXP[data$CROPDMGEXP!="b" & data$CROPDMGEXP!="m" & data$CROPDMGEXP!="k"] <- 0  
data$CROPDMGEXP[data$CROPDMGEXP=="b"] <- 1000000000  
data$CROPDMGEXP[data$CROPDMGEXP=="m"] <- 1000000  
data$CROPDMGEXP[data$CROPDMGEXP=="k"] <- 1000  
  
#Multiply DMG columns  
data$PROPDMG <- data$PROPDMG \* as.numeric(data$PROPDMGEXP)  
data$CROPDMG <- data$CROPDMG \* as.numeric(data$CROPDMGEXP)  
data <- select(data, EVTYPE, FATALITIES, INJURIES, PROPDMG, CROPDMG)

There are 985 levels of EVTYPE factor. We will try to reduce the number of unique values

data <- filter(data, !grepl("Summary", EVTYPE, ignore.case = TRUE))  
data$EVTYPE <- toupper(as.character(data$EVTYPE))  
data[grep("heat", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Heat"  
data[grep("warm", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Heat"  
data[grep("hot", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Heat"  
data[grep("HIGH TEMPER", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Heat"  
data[grep("HYPERTHERMIA", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Heat"  
data[grep("Record High", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Heat"  
  
data[grep("cold", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Cold"  
data[grep("cool", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Cold"  
data[grep("frost", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Cold"  
data[grep("HYPORTHERMIA", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Cold"  
data[grep("freez", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Cold"  
data[grep("frost", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Cold"  
data[grep("ice", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Cold"  
data[grep("icy", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Cold"  
data[grep("LOW TEMPER", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Cold"  
data[grep("RECORD LOW", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Cold"  
  
data[grep("tornado", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Tornado"  
data[grep("spout", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Tornado"  
  
data[grep("swell", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "High swells"  
data[grep("surf", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "High swells"  
data[grep("wave", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "High swells"  
data[grep("ROUGH SEAS", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "High swells"  
  
data[grep("wind", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "High wind/hurricane"  
data[grep("hurricane", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "High wind/hurricane"  
data[grep("wnd", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "High wind/hurricane"  
data[grep("TYPHOON", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "High wind/hurricane"  
data[grep("tsunami", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "High wind/hurricane"  
data[grep("storm", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "High wind/hurricane"  
  
data[grep("snow", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Heavy snow"  
data[grep("blizz", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Heavy snow"  
data[grep("wint", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Heavy snow"  
  
data[grep("flood", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Flood"  
data[grep("FLD", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Flood"  
data[grep("water", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Flood"  
data[grep("stream", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Flood"  
data[grep("tide", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Flood"  
  
data[grep("rain", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Heavy rain"  
data[grep("SHOWER", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Heavy rain"  
data[grep("precipit", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Heavy rain"  
  
data[grep("dry", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Drought"  
data[grep("drought", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Drought"  
  
data[grep("hail", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Hail"  
  
data[grep("fog", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Fog"  
  
data[grep("wet", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Wet"  
  
data[grep("fire", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Wildfire"  
  
data[grep("light", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Lightning"  
  
data[grep("rip", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Rip current"  
  
data[grep("avalanc", data$EVTYPE, ignore.case = TRUE), "EVTYPE"] <- "Avalanche"

All other values that a still in upper case we'll map to "Other" category. Then group all types together.

data$EVTYPE[data$EVTYPE == toupper(data$EVTYPE)] <- "Other"

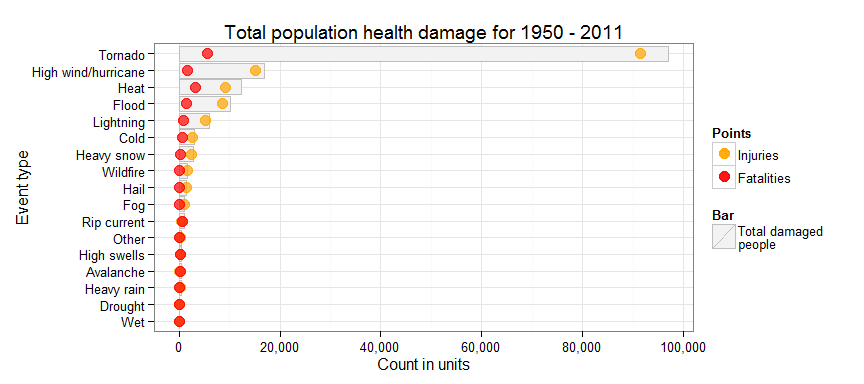
Group data and make columns with total harm mostly for ordering purposes:

data$EVTYPE <- factor(data$EVTYPE)  
data <- group\_by(data, EVTYPE) %>% summarise\_each(funs(sum))  
Hdata <- select(mutate(data, TOTALDMG = INJURIES + FATALITIES), EVTYPE, TOTALDMG, INJURIES, FATALITIES)  
Edata <- select(mutate(data, TOTALDMG = PROPDMG + CROPDMG), EVTYPE, TOTALDMG, PROPDMG, CROPDMG)  
rm(data)  
Edata <- arrange(Edata, desc(TOTALDMG))  
Hdata <- arrange(Hdata, desc(TOTALDMG))

## Results

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

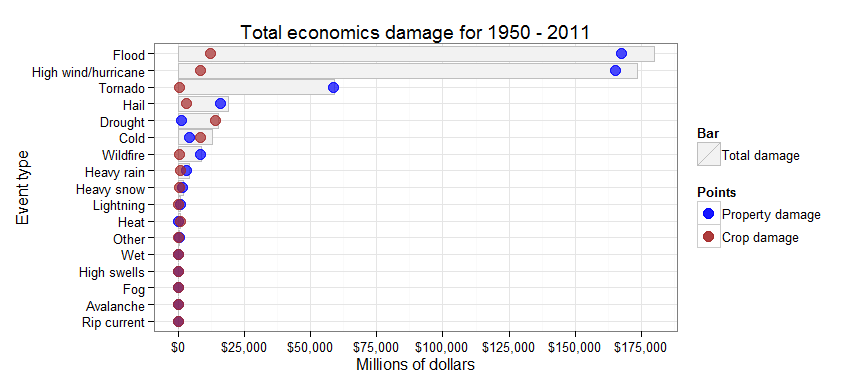
p <- ggplot(Hdata, aes(x = EVTYPE)) +  
 geom\_bar(aes(y = TOTALDMG, fill = "grey95"), stat="identity", colour = "grey") +  
 geom\_point(aes(y = INJURIES, colour = "orange"), alpha = 0.7, shape = 19, size = 4) +  
 geom\_point(aes(y = FATALITIES, colour = "red"), alpha = 0.7, shape = 19, size = 4) +  
 scale\_fill\_identity(name = 'Bar', guide = 'legend',labels = c('Total damaged\npeople')) +  
 scale\_colour\_manual(name = 'Points',   
 values =c('orange'='orange','red'='red'), labels = c('Injuries','Fatalities')) +  
 scale\_x\_discrete(limits = rev(Hdata$EVTYPE)) +  
 scale\_y\_continuous(labels = comma, breaks = seq(0, 100000, 20000)) +  
 ylab("Count in units") + xlab("Event type") +  
 ggtitle("Total population health damage for 1950 - 2011") +  
 theme\_bw() + coord\_flip()  
print(p)



Most damage to population health was done by **tornadoes**, **hurricanes**, **heat** and **floods**. Heat comes in second place if we'll count fatalities.

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

p <- ggplot(Edata, aes(x = EVTYPE)) +  
 geom\_bar(aes(y = TOTALDMG/1000000, fill = "grey95"), stat="identity", colour = "grey") +  
 geom\_point(aes(y = PROPDMG/1000000, colour = "blue"), alpha = 0.7, shape = 19, size = 4) +  
 geom\_point(aes(y = CROPDMG/1000000, colour = "brown"), alpha = 0.7, shape = 19, size = 4) +  
 scale\_fill\_identity(name = 'Bar', guide = 'legend',labels = c('Total damage')) +  
 scale\_colour\_manual(name = 'Points',   
 values =c('blue'='blue','brown'='brown'), labels = c('Property damage','Crop damage')) +  
 scale\_x\_discrete(limits = rev(Edata$EVTYPE)) +  
 scale\_y\_continuous(labels = dollar, breaks = seq(0, 175000, 25000)) +  
 ylab("Millions of dollars") + xlab("Event type") +  
 ggtitle("Total economics damage for 1950 - 2011") +  
 theme\_bw() + coord\_flip()  
print(p)



The to leaders are **floods** and **hurricanes**. If we are talking about crop, **drought** and **cold** did lots of harm too.