Understanding the Effects of Weather on Public Health and Economy

Lenny Fenster

Sunday, October 26, 2014

Synopsis

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, property damage, and crop damage. Preventing such outcomes to any extent possible is a key concern.

This project explored the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database. The weather events in the United States were observed against the number of injuries recorded for each weather event type as well as fatalities, property damage, and crop damage. It is shown that the most catastrophic type of weather event to population health both in terms of injury and fatality are tornadoes. However, the greatest damage from an economic perspective to both property and crops is Floods.

Data Processing

To ascertain which weather event types have the most harmful effects on both public health and the economy, the data from NOAA Storm Database was explored. This data exists in a comma-separated-value file compressed via the bzip2 algorithm to reduce its size. A check was made to ensure this file exists in the current directory and if not, downloads it from the appropriate location. Once we are assured the file exists, the data is read from it and stored in local variable. Functions for getting the data and reading it are used to encapsulate the logic needed for each of these steps.

```
##Make sure the raw data file exist.
##If data file does not yet exists, download the zip file and unzip the raw data file
getDataFiles <- function()
{
    ##check the subdirectories for train and test do not exist, extract from zip file
    if(!file.exists("StormData.csv.bz2"))
    {
        ## if raw zip file does not exist, download it
        if(!file.exists("StormData.csv.bz2"))
        {
            ## if doesnt exist download file
            download.file("https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2",des
        }
    }
}


readData <- function()
{
        stormData <- read.csv(stormDataFile<-bzfile("StormData.csv.bz2", "r"))
        close(stormDataFile)
        return (stormData)
}
```

```
getDataFiles()
stormData<-readData()
```

Data Cleansing

Exploratory data analysis followed downloaded and reading the data into a local dataset. Exploratory analysis showed that there was a great amount of variety in the verbiage describing the weather event types; to the extent of 898 unique descriptions. The contrasts with the documentation in *Section 2.1.1* on page 6 of the National Weather Service Storm Data Documentation that describes the 48 valid event types. The first order of data cleasing was to consolidate these 898 unique types in the 48 valid types as best as possible.

In an effort to best represent the intent behind the data entries and minimize any loss, I mapped a regular expression to each valid weather entry. For example, i inferred any entry with both the words *Dense* and *Smoke* in it to be a representation of the *Dense Smoke* weather event type and so I mapped *Dense Smoke* to the regular expression "(?=.*Dense)(?=.*Smoke)" to it. The regular expressions (the patterns) and the valid Event Types were both stored in the vectors permittedEventsPattern and permittedEventTypes respectively. Where there was the possibility of duplicate patterns returning the same entry (e.g., Flood and Flash Flood), the pattern was sorted from generic (e.g., Flood) to specific (e.g., Flash Flood) so only a more specific entry would be overwritten. The valid events and event patterns were collapsed together to form a mapping table named *goodEvTTable*.

```
#there are 898 unique event types in the raw dataset as determined by length(unique(toupper(stormData$E
uniqueEventTypes<-sort(unique(toupper(stormData$EVTYPE)))

#however, according to the storm data documentation only 48 specific event types are allowed
permittedEventsPattern<-c("^Astronomical Low", "^Avalanc", "^Blizzard", "(?=.*Cold|Wind)(?=.*Chill)",
                          "(?=.*Dense)(?=.*Fog)", "(?=.*Dense)(?=.*Smoke)", "(?=.*Drought)", "(?=.*Dust
                          "(?=.*Dust Storm)", "(?=.*Extreme)(?=.*Chill)", "(?=.*Flood)","(?=.*Forst|Fre
                          "(?=.*Funnel)","(?=.*Freezing Fog)", "(?=.*Hail)", "(?=.*Heat)", "(?=.*Heavy
                          "(?=.*High Surf)","(?=.*High Wind)", "(?=.*Hurricane|Typhoon)", "(?=.*Ice Sto
                          "(?=.*Lake)(?=.*Flood)","(?=.*Lightning)", "(?=.*Marine Hail)", "(?=.*Marine
                          "(?=.*Marine T)","(?=.*Rip Current)", "(?=.*Seiche)", "(?=.*Sleet)", "(?=.*St
                          "(?=.*Thunderstorm Wind)","(?=.*Tornado)", "(?=.*Tropical Depression)", "(?=.
                          "(?=.*Volcanic Ash)", "(?=.*Waterspout)", "(?=.*Wildfire)", "(?=.*Winter Storm
                          "(?=.*Coastal)(?=.*Flood)","(?=.*Debris)", "(?=.*Excessive Heat)","(?=.*Flash
                          )
permittedEvents<-c("Astronomical Low Tide","Avalanche","Blizzard","Cold/Wind Chill",
                   "Dense Fog","Dense Smoke","Drought","Dust Devil",
                   "Dust Storm","Extreme Cold/Wind Chill","Flood","Frost/Freeze",
                   "Funnel Cloud","Freezing Fog","Hail","Heat","Heavy Rain","Heavy Snow",
                   "High Surf","High Wind","Hurricane (Typhoon)","Ice Storm",
                   "Lake-Effect Snow","Lakeshore Flood","Lightning",
                   "Marine Hail","Marine High Wind",
                   "Marine Thunderstorm Wind","Rip Current","Seiche","Sleet","Storm Surge/Tide","Strong
                   "Thunderstorm Wind","Tornado","Tropical Depression","Tropical Storm","Tsunami",
                   "Volcanic Ash","Waterspout","Wildfire","Winter Storm","Winter Weather",
                   "Coastal Flood","Debris Flow","Excessive Heat","Flash Flood","Marine Strong Wind")

goodEvTTable<-data.frame(permittedEvents, permittedEventsPattern)
```

Additionally, columns which had no relevance to determining the impact on population health and economy were removed from investigation. A clean dataset was created that only included the first eight (primarily

demographic) variables as well as columns 23 through 28 which include the numbers for injuries, fatalities, property damage, crop damage, and exponential multipliers for both property damage and crop damage.

After removing unneeded columns, the mapping table was used to populate a new column with the mapped valid event type as determined by applying the mapped pattern to the EVTYPE variable. Any observations that did not match any of these patterns was omitted as invlid data due to an inability to infer a valid event type. The new variable, *GOODEVTYPE*, representing the valid event type according to the aforementioned table in the Storm Data Documentation is stored as a factor.

```
    #remove columns not needed for this evaluation
    cleanedStormData<-stormData[,c(1:8, 23:28)]
    cleanedStormData$GOODEVTYPE<-NA


    for (i in 1:length(permittedEvents)) {
        cleanedStormData[toupper(cleanedStormData$EVTYPE) %in% uniqueEventTypes[grep(goodEvTTable$permi
    }
    cleanedStormData<-na.omit(cleanedStormData)
    cleanedStormData$GOODEVTYPE<-as.factor(cleanedStormData$GOODEVTYPE)
```

Once the event types are cleansed, additional cleansing was performed to aide in the calculation for impact to the economy. The dataset includes variables that represent exponential multipliers to the base numbers for property damage and crop damage respectively. Inferring from the in *Section 2.7* on page 12 of the same document, "B|b" equates to billions, "M|m" to millions, "K|k" to thousands, and following that logic "H|h" would equate to hundreds. Thus, these characters were replaced with the proper exponential value representing each of them (i.e., 9, 6, 3, and 2). There were additional characters (namely, empty string, plus, minus, and question mark (‘’,’+‘,’-‘,’?’)) for which an exponential value could not be determined. The column was transformed to a numeric and any non-numeric values, which were then represented as NA, were converted to zero. An additional column, *TOTALDMG*, representing the total economic damage due to a weather event, was created and populated with the (value from PROPDMG * 10ˆvalue from PROPDMGEXP) + (value from CROPDMG * 10ˆvalue from CROPDMGEXP)

```
    #convert exponential 'keys' to numerics
    expkeys<-c("h","k","m","b")
    expvalue<-c(2,3,6,9)
    for(i in 1:length(expkeys)) {
        cleanedStormData$PROPDMGEXP<-gsub(expkeys[i],expvalue[i],cleanedStormData$PROPDMGEXP, ignore.cas
        cleanedStormData$CROPDMGEXP<-gsub(expkeys[i],expvalue[i],cleanedStormData$CROPDMGEXP, ignore.cas
    }

    #intentionally convert all non-numeric characters that are left (e.g., " +/-/?") to NA and then zer
    suppressWarnings(cleanedStormData$PROPDMGEXP<-as.numeric(cleanedStormData$PROPDMGEXP))
    suppressWarnings(cleanedStormData$CROPDMGEXP<-as.numeric(cleanedStormData$CROPDMGEXP))
    cleanedStormData$PROPDMGEXP[is.na(cleanedStormData$PROPDMGEXP)]<-0
    cleanedStormData$CROPDMGEXP[is.na(cleanedStormData$CROPDMGEXP)]<-0

    cleanedStormData$TOTALDMG = (cleanedStormData$PROPDMG * 10^cleanedStormData$PROPDMGEXP) +
                                (cleanedStormData$CROPDMG * 10^cleanedStormData$CROPDMGEXP)
```

Finally, a concise "summation table" was created that held an observation for every weather event type and the variables for the sum of injuries, fatalities, and total economic damage. This functionality was encapsulated in a function named *createSummationTable*.

```
createSummationTable<-function(stormData)
{
    aggregatedStormData<-aggregate(stormData$INJURIES, list(EVENTTYPE=stormData$GOODEVTYPE), sum)
    fatalities<-aggregate(stormData$FATALITIES, list(EVENTTYPE=stormData$GOODEVTYPE), sum)
    totaldmg<-aggregate(stormData$TOTALDMG, list(EVENTTYPE=stormData$GOODEVTYPE), sum)
    colnames(aggregatedStormData)[2]<-"injuries"
    colnames(fatalities)[2]<-"fatalities"
    colnames(totaldmg)[2]<-"totaldmg"
    aggregatedStormData<-merge(aggregatedStormData, merge(fatalities, totaldmg))

    return (aggregatedStormData)
}

aggregateStormData<-createSummationTable(cleanedStormData)
```

Results

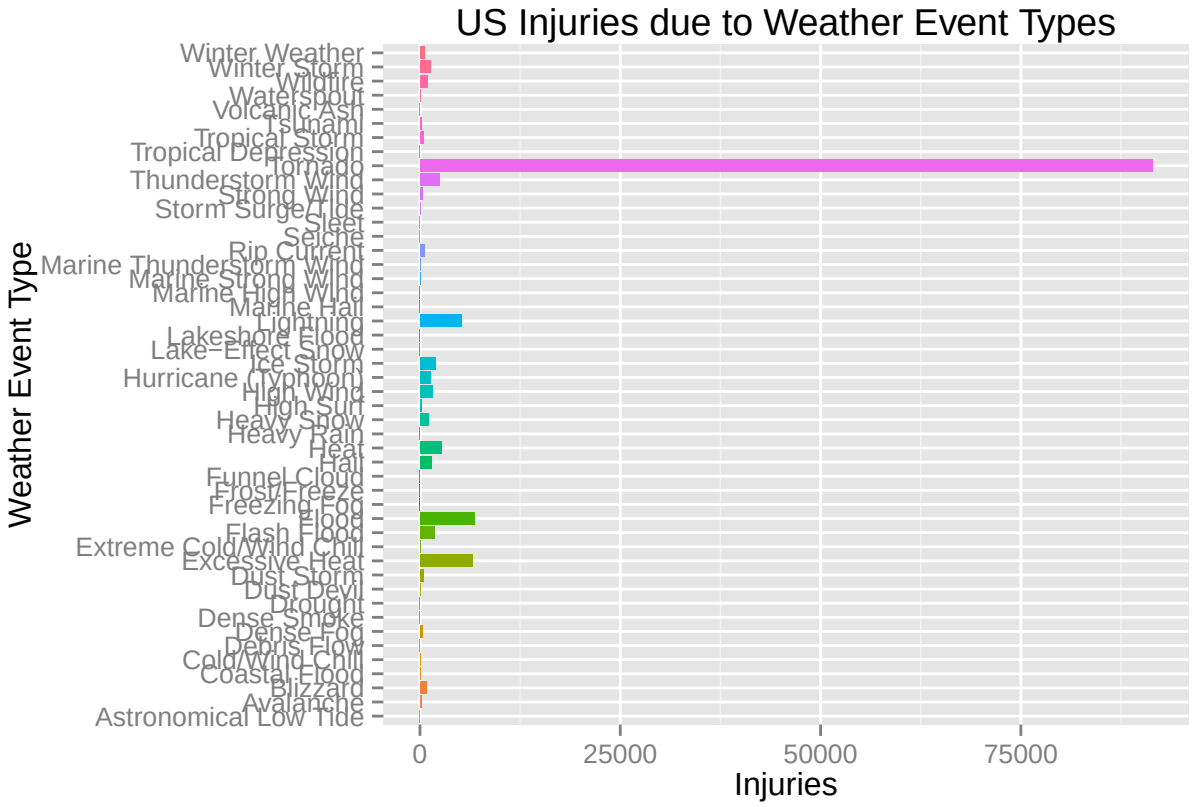
With the summation table created, determining the impact that each weather event as on population health and the economy becomes straightforward.

Effect of Weather Events on Injuries

The effect of weather events on injuries can be observed by plotting the EVENTTYPE against the observed injuries as shown below.

```
library(ggplot2)
    injuryplot<-ggplot(aggregateStormData, aes(x = EVENTTYPE, y = injuries, fill = EVENTTYPE)) +
        geom_bar(stat = "identity") +
        xlab("Weather Event Type") +
        ylab("Injuries") +
        guides(fill=FALSE) +
        ggtitle("US Injuries due to Weather Event Types") +
        coord_flip()

        print(injuryplot)
```



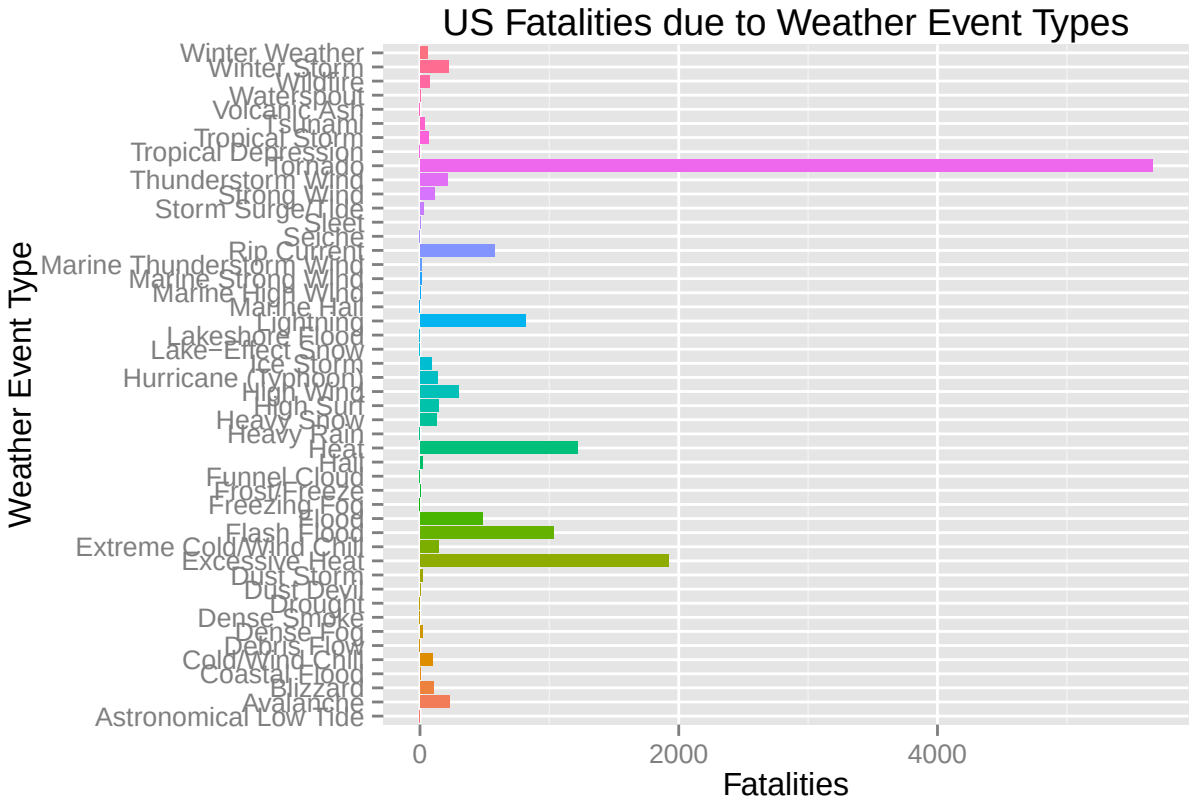
We can observe in this plot that Tornado is the weather event type that is associated with the greatest number of injuries at 9.1364×104.

Effect of Weather Events on Fatalities

Injuries are just one of the measurements related to public health concerns. Fatalities are the other. The effect of weather events on fatalities can be observed by plotting the EVENTTYPE against the observed fatalities as shown below.

```
    fatalplot<-ggplot(aggregateStormData, aes(x = EVENTTYPE, y = fatalities, fill = EVENTTYPE)) +
        geom_bar(stat = "identity") +
        xlab("Weather Event Type") +
        ylab("Fatalities") +
        guides(fill=FALSE) +
        ggtitle("US Fatalities due to Weather Event Types") +
        coord_flip()

        print(fatalplot)
```



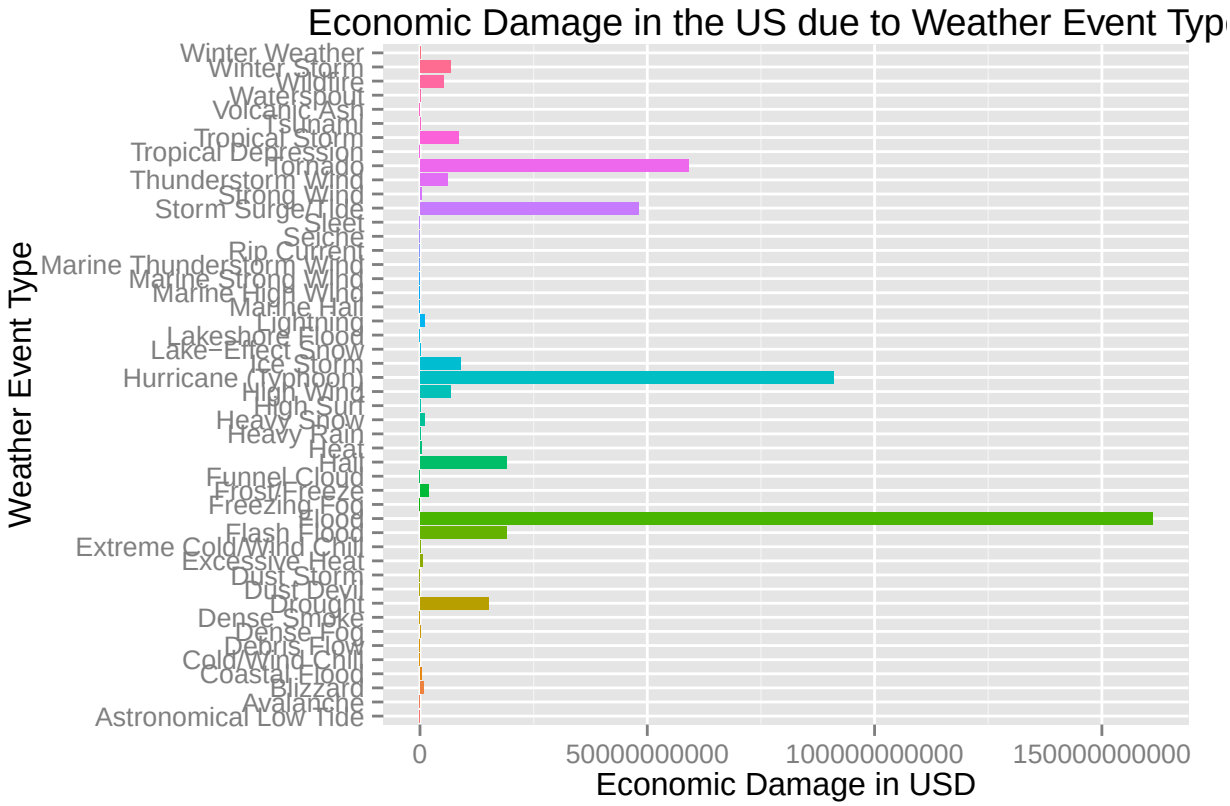
We can observe in this plot that Tornado is the weather event type that is associated with the greatest number of fatalities at 5658.

Effect of Weather Events on the Economy

Lastly, the effect of weather events on the economy, namely property and crop damage, can be observed by plotting the EVENTTYPE against the observed total damage (totaldmg field) as shown below.

```
    options(scipen=12)
    economicplot<-ggplot(aggregateStormData, aes(x = EVENTTYPE, y = totaldmg, fill = EVENTTYPE)) +
        geom_bar(stat = "identity") +
        xlab("Weather Event Type") +
        ylab("Economic Damage in USD") +
        guides(fill=FALSE) +
        ggtitle("Economic Damage in the US due to Weather Event Types") +
        coord_flip()

        print(economicplot)
```



We can observe in this plot that Flood is the weather event type that is associated with the greatest impact on the economy at 161023285629.