Impact of Extreme Weather Events in USA

Author: Ranish Bera

Synopsis

The goal of the assignment is to explore the NOAA Storm Database and explore the effects of severe weather events on both population and economy.

The database covers the time period between 1950 and November 2011. In the earlier years of the database there are generally fewer events recorded, most likely due to a lack of good records. More recent years should be considered more complete.

The analysis aims to investigate which different types of severe weather events are most harmful on the population health in respect of general injuries and fatalities. Further the economic consequences will be analyzed by exploring the financial damage done to both general property and agriculture (i.e. crops).

Data Analysis

The raw data is available in coursera course website link. The documentation of the data is available here.

We download the data in default directory and load the data in R.

```
> url<-"https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
> download.file(url,destfile = "FStormData.csv")
trying URL 'https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2'
Content type 'application/bzip2' length 49177144 bytes (46.9 MB)
downloaded 46.9 MB

> data<-read.csv("FStormData.csv")
```

The URL for the data was copied from the link provided in Coursera course. We perform a few exploratory analyses to understand the data. We supplement our knowledge by reading the documentation provided.

```
> str(data)
'data.frame':   902297 obs. of  37 variables:
 $ STATE__   : num  1 1 1 1 1 1 1 1 1 1 ...
 $ BGN_DATE  : Factor w/ 16335 levels "1/1/1966 0:00:00",..: 6523 6523 4242 11116 2224 2224 2260 383 3980 3980 ...
 $ BGN_TIME  : Factor w/ 3608 levels "00:00:00 AM",..: 272 287 2705 1683 2584 3186 242 1683 3186 3186 ...
 $ TIME_ZONE : Factor w/ 22 levels "ADT","AKS","AST",..: 7 7 7 7 7 7 7 7 7 7 ...
 $ COUNTY    : num  97 3 57 89 43 77 9 123 125 57 ...
 $ COUNTYNAME: Factor w/ 29601 levels "","5NM E OF MACKINAC BRIDGE TO PRESQUE ISLE LT MI",..: 13513 1873 4598 10592 4372 10094 1973 23873 24418 4598 ...
 $ STATE     : Factor w/ 72 levels "AK","AL","AM",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ EVTYPE    : Factor w/ 985 levels "   HIGH SURF ADVISORY",..: 834 834 834 834 834 834 834 834 834 834 ...
 $ BGN_RANGE : num  0 0 0 0 0 0 0 0 0 0 ...
 $ BGN_AZI   : Factor w/ 35 levels "","  N"," NW",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ BGN_LOCATI: Factor w/ 54429 levels "","- 1 N Albion",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ END_DATE  : Factor w/ 6663 levels "","1/1/1993 0:00:00",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ END_TIME  : Factor w/ 3647 levels ""," 0900CST",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ 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   : Factor w/ 24 levels "","E","ENE","ESE",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ END_LOCATI: Factor w/ 34506 levels "","- .5 NNW",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ 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: Factor w/ 19 levels "","-","?","+",..: 17 17 17 17 17 17 17 17 17 17 ...
 $ CROPDMG   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ CROPDMGEXP: Factor w/ 9 levels "","?","0","2",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ WFO       : Factor w/ 542 levels ""," CI","$AC",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ STATEOFFIC: Factor w/ 250 levels "","ALABAMA, Central",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ ZONENAMES : Factor w/ 25112 levels "","
 $ 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   : Factor w/ 436781 levels "","-2 at Deer Park\n",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
$ REMARKS    : Factor w/ 436781 levels "","-2 at Deer Park\n",..: 1 1 1 1 1 1 1 1 1 1 ...
$ REFNUM     : num  1 2 3 4 5 6 7 8 9 10 ...
```

Since the aim of the analysis is to answer the following questions:-

1. Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health?
2. Across the United States, which types of events have the greatest economic consequences?

Many of the variables appearing are not relevant to the analysis. Specifically, the following variables are of interest:-

- EVTYPE
- FATALITIES
- INJURIES
- PROPDMG
- PROPDMGEXP
- CROPDMG
- CROPDMGEXP

The description of these variables is available in documentation. These variables are required because we are interested only in the impact of severe weather in entire USA, not in any specific state. The variables FATALITIES and INJURIES provide the data on direct deaths and injuries because of the severe weather event. These variables are relevant to calculate the damage to population health. Similarly, the variables – PROPDMG, PROPDMGEXP, CROPDMG, CROPDMGEXP are relevant to calculate the extent of economical impact the severe weather event had made.

Hence, we create the subset data to do further analysis on the data of our interest.

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

To do further analysis, we need the following libraries loaded in R:-

- library(plyr)
- library(ggplot2)
- library(gridExtra)
- library(grid)

Impact on Population Health

The fatalities as well as the injuries are summarized according to the event type and then, respectively, ordered decreasingly. This will help us find out the event that caused maximum fatalities and injuries.

```
> health<-ddply(sub_dat, .(EVTYPE), summarize,fatalities = sum(FATALITIES),injuries = sum(INJURIES))
> fatal<-health[order(health$fatalities, decreasing = T), ]
> injuries<-health[order(health$injuries, decreasing = T), ]
```

Impact on Economy

Since the exponential values are stored in a separate column describing their value with letters (h = hundred, k = thousand, m = million, b = billion), the calculation of the financial damage turns out to be slightly tricky. In a first step a function that converts the letter value of the exponent to a usable number must be implemented.

```
> getExp <- function(e) {
+     if (e %in% c("h", "H"))
+         return(2)
+     else if (e %in% c("k", "K"))
+         return(3)
+     else if (e %in% c("m", "M"))
+         return(6)
+     else if (e %in% c("b", "B"))
+         return(9)
+     else if (!is.na(as.numeric(e)))
+         return(as.numeric(e))
+     else if (e %in% c("", "-", "?", "+"))
+         return(0)
+     else {
+         stop("Invalid value.")
+     }
+ }
```

The economic impact, property damage and crop damage, is calculated and stored in another dataset. Then the financial impact of property and crop damage is summarized according to event type. We clean the dataset of any entry that did not cause any economic damage. The data is then sorted.

```
> propExp <- sapply(sub_dat$PROPDMGEXP, FUN=getExp)
> sub_dat$propDamage <- sub_dat$PROPDMG * (10 ** propExp)
> cropExp <- sapply(sub_dat$CROPDMGEXP, FUN=getExp)
> sub_dat$cropDamage <- sub_dat$CROPDMG * (10 ** cropExp)
> econDamage <- ddply(sub_dat, .(EVTYPE), summarize,propDamage = sum(propDamage), cropDamage = sum(cropDamage))
> econDamage <- econDamage[(econDamage$propDamage > 0 || econDamage$cropDamage > 0), ]
> propDmgSorted <- econDamage[order(econDamage$propDamage, decreasing = T), ]
> cropDmgSorted <- econDamage[order(econDamage$cropDamage, decreasing = T), ]
```

The data has been processed properly.

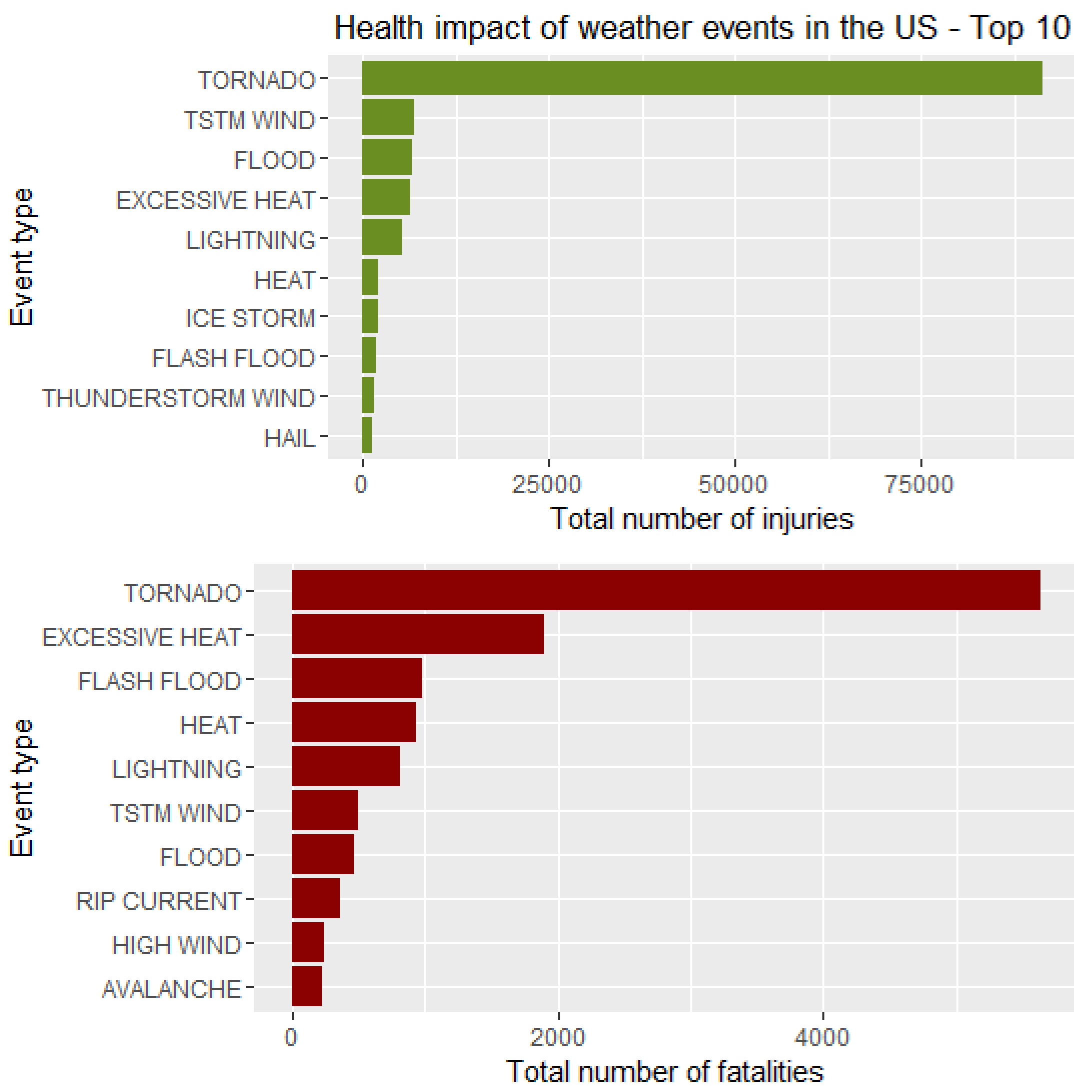
Results

Impact on Population Health

Lists of the Top 5 weather events affecting the population's health (injuries and deaths) are shown. For injuries as well as fatal events the most devastating events are tornados in the given time period. We also plotted the fatalities and injuries against the events (top 10). The top contributor to fatalities as well as injuries is "Tornado".

```
> head(fatal[, c("EVTYPE", "fatalities")],5)
            EVTYPE fatalities
834        TORNADO       5633
130 EXCESSIVE HEAT       1903
153    FLASH FLOOD        978
275           HEAT        937
464      LIGHTNING        816
>
> head(injuries[, c("EVTYPE", "injuries")],5)
            EVTYPE injuries
834        TORNADO    91346
856      TSTM WIND     6957
170          FLOOD     6789
130 EXCESSIVE HEAT     6525
464      LIGHTNING     5230
```

```
> p1 <- ggplot(data=head(injury,10), aes(x=reorder(EVTYPE, injuries), y=injuries)) +
+    geom_bar(fill="olivedrab",stat="identity")  + coord_flip() +
+     ylab("Total number of injuries") + xlab("Event type") +
+     ggtitle("Health impact of weather events in the US - Top 10") +
+     theme(legend.position="none")
> p2 <- ggplot(data=head(fatal,10), aes(x=reorder(EVTYPE, fatalities), y=fatalities)) +
+     geom_bar(fill="red4",stat="identity") + coord_flip() +
+     ylab("Total number of fatalities") + xlab("Event type") +
+     theme(legend.position="none")
> grid.arrange(p1, p2, nrow =2)
```



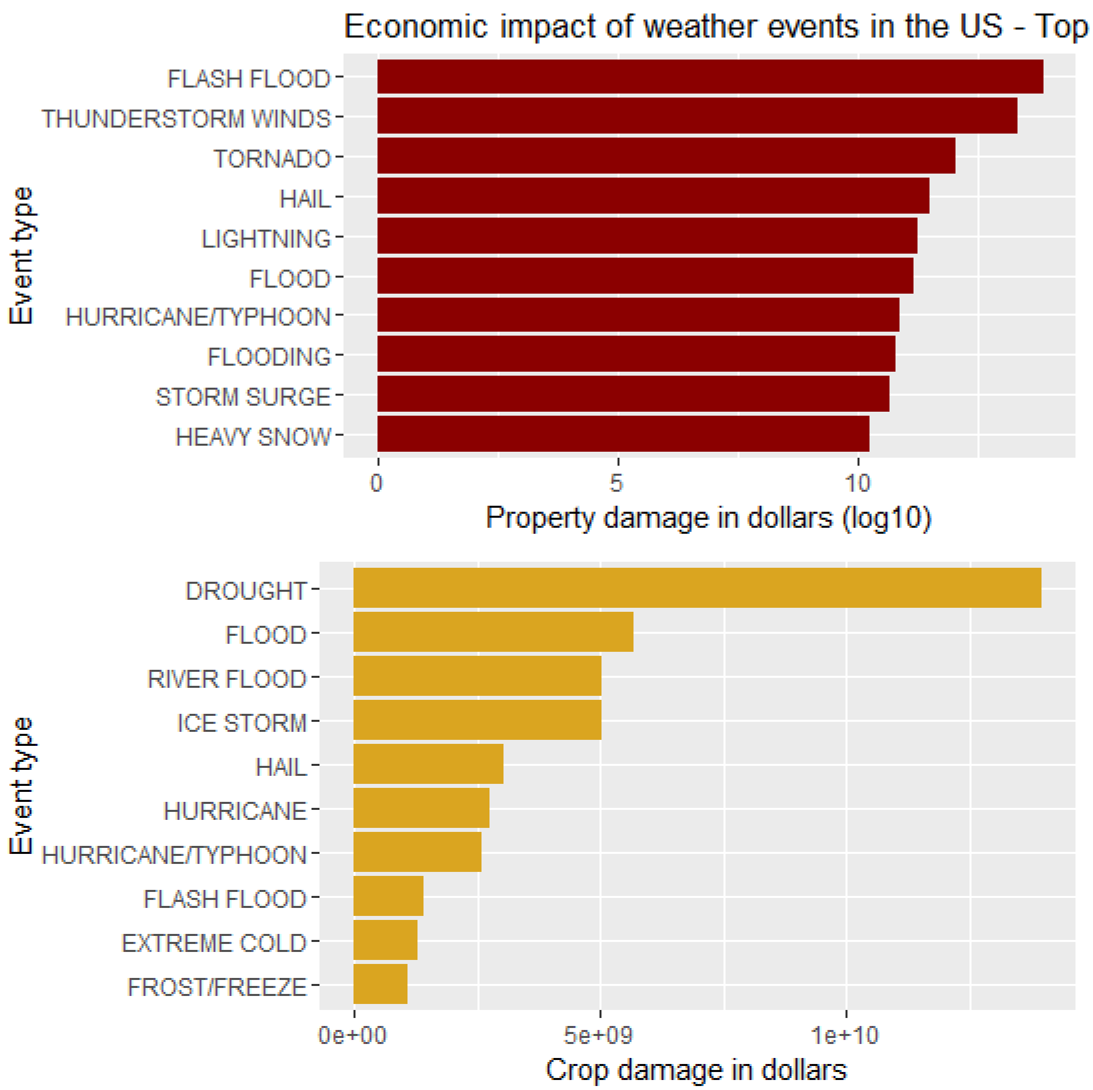
The next most damaging (in terms of health) are Excessive Heat and Flash Flood.

Impact on Economy

Lists of the Top 5 weather events causing financial damage to both property and crops are shown below.

```
> head(propDmgSorted[, c("EVTYPE", "propDamage")], 5)
                 EVTYPE   propDamage
153         FLASH FLOOD 6.820237e+13
786 THUNDERSTORM WINDS 2.086532e+13
834             TORNADO 1.078951e+12
244                HAIL 3.157558e+11
464           LIGHTNING 1.729433e+11
> head(cropDmgSorted[, c("EVTYPE", "cropDamage")], 5)
         EVTYPE  cropDamage
95      DROUGHT 13972566000
170       FLOOD  5661968450
590 RIVER FLOOD  5029459000
427   ICE STORM  5022113500
244        HAIL  3025974480
```

```
> p1 <- ggplot(data=head(propDmgSorted,10), aes(x=reorder(EVTYPE, propDamage), y=log10(propDamage), fill=propDamage )) +
+     geom_bar(fill="darkred", stat="identity") + coord_flip() +
+     xlab("Event type") + ylab("Property damage in dollars (log10)") +
+     ggtitle("Economic impact of weather events in the US - Top 10") +
+     theme(plot.title = element_text(hjust = 0))
>
> p2 <- ggplot(data=head(cropDmgSorted,10), aes(x=reorder(EVTYPE, cropDamage), y=cropDamage, fill=cropDamage)) +
+     geom_bar(fill="goldenrod", stat="identity") + coord_flip() +
+     xlab("Event type") + ylab("Crop damage in dollars") +
+     theme(legend.position="none")
>
> grid.arrange(p1, p2, ncol=1, nrow =2)
>
```



For property flash floods, thunderstorms and tornados seem to cause by far the most damage (orders of magnitude above the other event types!). Probably little surprising is that the weather event causing most financial damage in respect to agriculture (i.e. crops) is drought. Noteworthy damage is also done by flood events as well as hail and ice storms. The findings described above become even more evident when we plot the Top 10 events for both property and crop damage! (Note that for plotting the property damage a logarithmic scale was used in order to increase the readability of the plot).

Appendix - R Code

```
url<-"https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"

download.file(url,destfile = "FStormData.csv")

data<-read.csv("FStormData.csv")

str(data)

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


#Loading libraries required

library(plyr)

library(ggplot2)

library(gridExtra)

library(grid)


#Crating table for analysis

health<-ddply(sub_dat, .(EVTYPE), summarize,fatalities = sum(FATALITIES),injuries = sum(INJURIES))

fatal<-health[order(health$fatalities, decreasing = T), ]#Ordered data for fatalities

injuries<-health[order(health$injuries, decreasing = T), ]#Ordered data for injuries
```

```r
#Function for converting denotations into power of 10
getExp <- function(e) {
  if (e %in% c("h", "H"))
    return(2)
  else if (e %in% c("k", "K"))
    return(3)
  else if (e %in% c("m", "M"))
    return(6)
  else if (e %in% c("b", "B"))
    return(9)
  else if (!is.na(as.numeric(e)))
    return(as.numeric(e))
  else if (e %in% c("", "-", "?", "+"))
    return(0)
  else {
    stop("Invalid value.")
  }
}
```

#Getting data for economic damage

propExp <- sapply(sub_dat$PROPDMGEXP, FUN=getExp)

sub_dat$propDamage <- sub_dat$PROPDMG * (10 ** propExp)

cropExp <- sapply(sub_dat$CROPDMGEXP, FUN=getExp)

sub_dat$cropDamage <- sub_dat$CROPDMG * (10 ** cropExp)

econDamage <- ddply(sub_dat, .(EVTYPE), summarize,propDamage = sum(propDamage), cropDamage = sum(cropDamage))

econDamage <- econDamage[(econDamage$propDamage > 0 || econDamage$cropDamage > 0),]%Those that do not cause financial loss will be ommitted

propDmgSorted <- econDamage[order(econDamage$propDamage, decreasing = T),]%Ordered data for property damage

cropDmgSorted <- econDamage[order(econDamage$cropDamage, decreasing = T),]%Ordered data for crop damage

#Results

#Most Injury Event

head(fatal[, c("EVTYPE", "fatalities")],5)

#Most Injury Event

head(injuries[, c("EVTYPE", "injuries")],5)

```
#Graphic reprsentation

p1 <- ggplot(data=head(injury,10), aes(x=reorder(EVTYPE, injuries), y=injuries)) +
   geom_bar(fill="olivedrab",stat="identity")  + coord_flip() +
    ylab("Total number of injuries") + xlab("Event type") +
    ggtitle("Health impact of weather events in the US - Top 10") +
    theme(legend.position="none")

p2 <- ggplot(data=head(fatal,10), aes(x=reorder(EVTYPE, fatalities), y=fatalities)) +
    geom_bar(fill="red4",stat="identity") + coord_flip() +
    ylab("Total number of fatalities") + xlab("Event type") +
    theme(legend.position="none")

grid.arrange(p1, p2, nrow =2)

#Economic Damage results

#Property damage
head(propDmgSorted[, c("EVTYPE", "propDamage")], 5)

#Crop Damage

head(cropDmgSorted[, c("EVTYPE", "cropDamage")], 5)
```

```
#Plot

p1 <- ggplot(data=head(propDmgSorted,10), aes(x=reorder(EVTYPE, propDamage),
y=log10(propDamage), fill=propDamage )) +
   geom_bar(fill="darkred", stat="identity") + coord_flip() +
   xlab("Event type") + ylab("Property damage in dollars (log10)") +
   ggtitle("Economic impact of weather events in the US - Top 10") +
   theme(plot.title = element_text(hjust = 0))


p2 <- ggplot(data=head(cropDmgSorted,10), aes(x=reorder(EVTYPE, cropDamage), y=cropDamage,
fill=cropDamage)) +
   geom_bar(fill="goldenrod", stat="identity") + coord_flip() +
   xlab("Event type") + ylab("Crop damage in dollars") +
   theme(legend.position="none")


grid.arrange(p1, p2, ncol=1, nrow =2)
```