Storm Data Analysis - Finding the worst type of Storms

SiddharthIITG

21 October 2015

Synopsis

In this analysis storm data from NOAA Storm Database has been used to answer two basic questions about severe weather events:

- 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?

```
library("ggplot2")
library("plyr")
library("tm")
library("gridExtra")
```

Data Processing

This section deals with reading in the data and processing it to do the analysis.

```
data = read.csv("./repdata-data-StormData.csv.bz2")
```

Programmatically extracting the reference categories from the National Weather Service Instruction 10-1605 PDF document.

```
pdf <- readPDF(control = list(c(text = "-layout")))
pdf <- pdf(elem=list(uri="nws_i10_1605.pdf"),language="en")
keep = c(pdf$content[seq(397, 420)], pdf$content[seq(425, 448)])
keep
```

```
##  [1] "Astronomical Low Tide"    "Avalanche"
##  [3] "Blizzard"                 "Coastal Flood"
##  [5] "Cold/Wind Chill"          "Debris Flow"
##  [7] "Dense Fog"                "Dense Smoke"
##  [9] "Drought"                  "Dust Devil"
## [11] "Dust Storm"               "Excessive Heat"
## [13] "Extreme Cold/Wind Chill"  "Flash Flood"
## [15] "Flood"                    "Frost/Freeze"
## [17] "Funnel Cloud"             "Freezing Fog"
## [19] "Hail"                     "Heat"
## [21] "Heavy Rain"               "Heavy Snow"
## [23] "High Surf"                "High Wind"
## [25] "Hurricane (Typhoon)"      "Ice Storm"
## [27] "Lake-Effect Snow"         "Lakeshore Flood"
## [29] "Lightning"                "Marine Hail"
```

```
## [31] "Marine High Wind"         "Marine Strong Wind"
## [33] "Marine Thunderstorm Wind" "Rip Current"
## [35] "Seiche"                   "Sleet"
## [37] "Storm Surge/Tide"         "Strong Wind"
## [39] "Thunderstorm Wind"        "Tornado"
## [41] "Tropical Depression"      "Tropical Storm"
## [43] "Tsunami"                  "Volcanic Ash"
## [45] "Waterspout"               "Wildfire"
## [47] "Winter Storm"             "Winter Weather"
```

The EVTYPE Variable consists of over 900 factor variables. Most of this is because of spelling mistake and shorthands being used. Most of these occur rarely and can be weeded out without any difference to the final result.

Extracting the 50 most frequent EVTYPE's from the data.

```
freq  = count(data, 'EVTYPE')
freq = arrange(freq, desc(freq))
freq = freq[1:50,]
head(freq)
```

```
##              EVTYPE   freq
## 1              HAIL 288661
## 2         TSTM WIND 219940
## 3 THUNDERSTORM WIND  82563
## 4           TORNADO  60652
## 5       FLASH FLOOD  54277
## 6             FLOOD  25326
```

The most frequent storm type is **Hail Storm**.

```
detach("package:plyr", unload=TRUE)
library("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
```

Subsetting the data to rows which have EVTYPE in freq.

Also, removing the rows where all of property damage, crop damage and fatalities are zero. This is beacause any storm event that does not do any damage is of no use to us in suggesting the most dangerous storm event.

```
df = merge(data, freq, by = "EVTYPE")
df = filter(df, FATALITIES != 0 | PROPDMG != 0 | CROPDMG != 0)
```

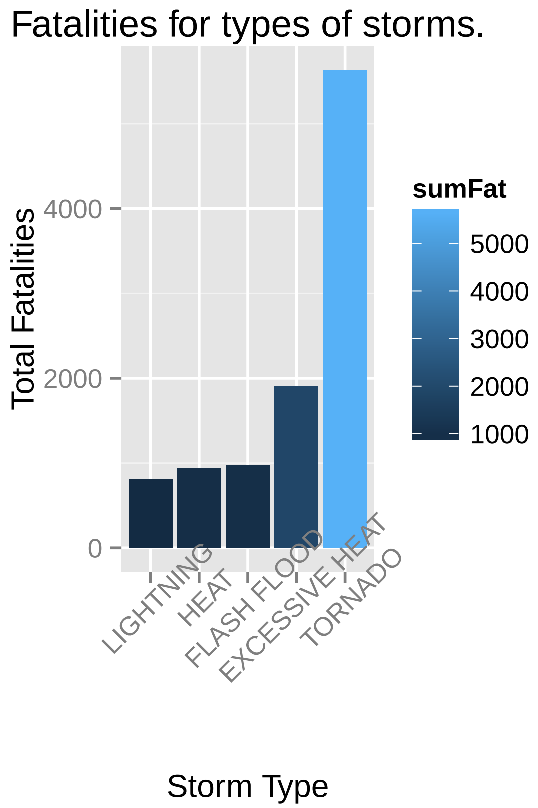
Data Analysis

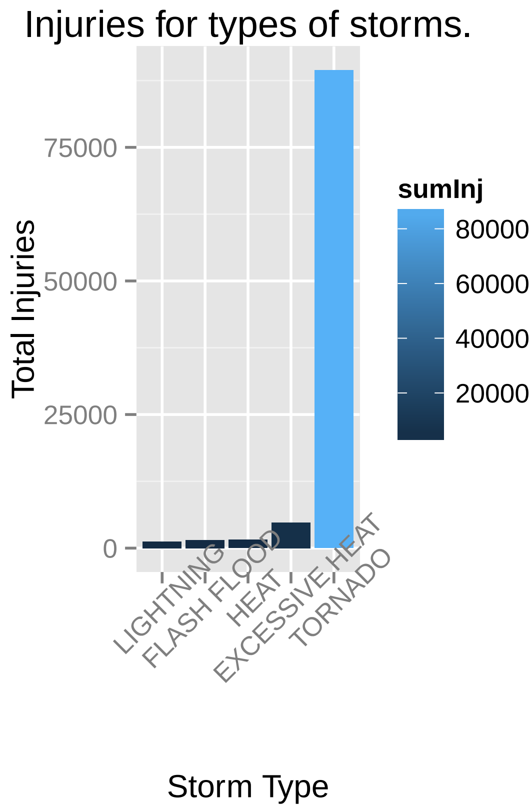
In this section graphs and tables have been made for the analysis. Specifically, the following graphs are plotted:

- Total fatalities for types of storms
- Total injuries for types of storms
- Total damage for types of storms

Damage to population health

```
grouping <- group_by(df, EVTYPE)
sumHealth = summarize(grouping,sumFat = sum(FATALITIES),
                      sumInj =  sum(INJURIES))
sumHealth = arrange(sumHealth, desc(sumFat))

sumHealth = sumHealth[1:5,]
plot1 = ggplot(sumHealth, aes(reorder(factor(EVTYPE), sumFat),
                       y = sumFat, fill = sumFat)) +
  geom_bar(stat = "identity") +
  labs(title = "Fatalities for types of storms.",
       x = "Storm Type", y = "Total Fatalities") +
        theme(axis.text.x = element_text(angle=45))

sumHealth = arrange(sumHealth, desc(sumInj))

sumHealth = sumHealth[1:5,]
plot2 = ggplot(sumHealth, aes(reorder(factor(EVTYPE), sumInj),
                               y = sumInj, fill = sumInj)) +
  geom_bar(stat = "identity") +
  labs(title = "Injuries for types of storms.",
       x = "Storm Type", y = "Total Injuries") +
  theme(axis.text.x = element_text(angle=45))

grid.arrange(plot1, plot2, ncol=2)
```





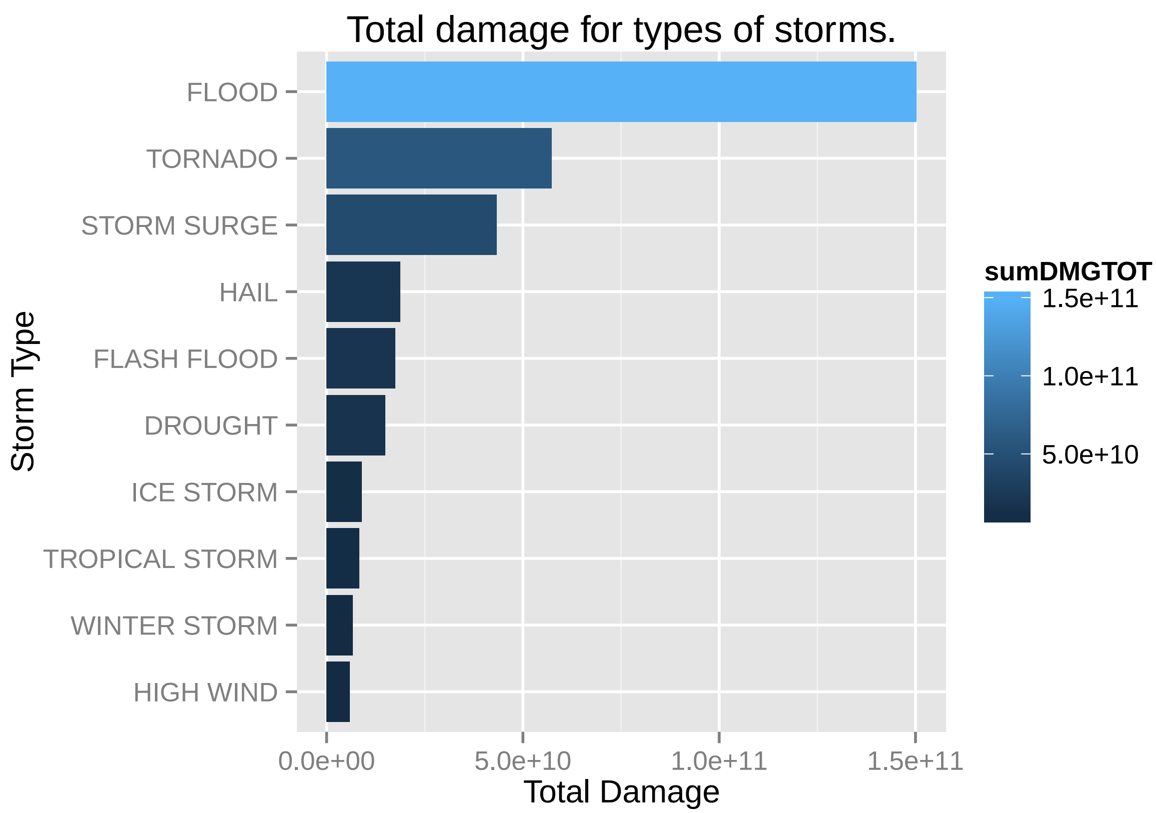
It is clear from the above graphs that **tornados** have been the worst affecting storm type for the **health of the population**. The fatalities and the injuries caused by tornados far outnumber that of the other storm types.

Economic Damage

In this section the damage to economy has been studied.

```
df$PROPMUL = 0 # Defining a new variable for obtaining the value of the PROPDMGEXP variable
df$CROPMUL = 0 # Defining a new variable for obtaining the value of the CROPDMGEXP variable

df[as.character(df$CROPDMGEXP) == "K",]$CROPMUL = 1000
df[as.character(df$CROPDMGEXP) == "M",]$CROPMUL = 1000000
df[as.character(df$CROPDMGEXP) == "B",]$CROPMUL = 1000000000

df[as.character(df$PROPDMGEXP) == "B",]$PROPMUL = 1000000000
df[as.character(df$PROPDMGEXP) == "M",]$PROPMUL = 1000000
df[as.character(df$PROPDMGEXP) == "K",]$PROPMUL = 1000


df$PROPDMGTOT = df$PROPDMG*df$PROPMUL # property damage for a  storm
df$CROPDMGTOT = df$CROPDMG*df$CROPMUL # crop damage for a storm

grouping <- group_by(df, EVTYPE)
sumDMG = summarize(grouping,sumPROPDMG = sum(PROPDMGTOT),
```

```
                       sumCROPDMG =  sum(CROPDMGTOT))

sumDMG$sumDMGTOT = sumDMG$sumPROPDMG + sumDMG$sumCROPDMG #total economic damage
sumDMG = arrange(sumDMG, desc(sumDMGTOT))
sumDMGhead = sumDMG[1:10,]

 plot = ggplot(sumDMGhead, aes(reorder(factor(EVTYPE), sumDMGTOT),
                         y = sumDMGTOT, fill = sumDMGTOT)) +
     geom_bar(stat = "identity") + coord_flip() +
   labs(title = "Total damage for types of storms.",
        x = "Storm Type", y = "Total Damage")
 plot
```



We can see from the above figure that **floods** cause the highest amount of total economic damage.

Let us now see which storm types cause the highest amount of damage to property and crop independently.

Damage to Property

```
sumDMG = arrange(sumDMG, desc(sumPROPDMG))
sumDMGhead = sumDMG[1:5,]
sumDMGhead
```

```
## Source: local data frame [5 x 4]
```

```
##
##        EVTYPE   sumPROPDMG sumCROPDMG    sumDMGTOT
##        (fctr)        (dbl)      (dbl)        (dbl)
## 1       FLOOD 144657709800 5661968450 150319678250
## 2     TORNADO  56925660480  414953110  57340613590
## 3 STORM SURGE  43323536000       5000  43323541000
## 4 FLASH FLOOD  16140811510 1421317100  17562128610
## 5        HAIL  15727366720 3025537450  18752904170
```

We can see that **floods** have caused the highest amount of damage to **property**.

Damage to Crop

```
sumDMG = arrange(sumDMG, desc(sumCROPDMG))
sumDMGhead = sumDMG[1:5,]
sumDMGhead
```

```
## Source: local data frame [5 x 4]
##
##        EVTYPE   sumPROPDMG  sumCROPDMG    sumDMGTOT
##        (fctr)        (dbl)       (dbl)        (dbl)
## 1     DROUGHT   1046106000 13972566000  15018672000
## 2       FLOOD 144657709800  5661968450 150319678250
## 3   ICE STORM   3944927810  5022113500   8967041310
## 4        HAIL  15727366720  3025537450  18752904170
## 5 FLASH FLOOD  16140811510  1421317100  17562128610
```

We can see that **droughts** have caused the highest amount of damage to **crop**.

Results

Looking at the graphs the following points can be ascertained:

- **Floods** have caused the highest amount of damage to the economy since the time this data has been collected.
- **Droughts** have caused the most amount of damage to the Crops.
- **Tornados** have had the worst impact on the health of the population.