Class 5 - Peer Assessment 2: Analysis of Effects of Severe Weather Events

Nicholas Ng

Sunday, January 25, 2015

Abstract

This report is produced as part of the requirements of the 2nd Peer Assessment in the Reproducible Research class of the Data Science Specialisation. Based on the data obtained from the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database, this report identifies the types of weather events that have the greatest effect on: 1. Human health, and 2. Economic damage

Data Processing

Before any transformations can be done on the data, the required data processing packages need to be loaded.

```
if (!suppressMessages(require("dplyr"))){
    install.packages("dplyr")
    suppressMessages(require("dplyr"))
}

if (!suppressMessages(require("data.table"))){
    install.packages("data.table")
    suppressMessages(require("data.table"))
}
```

As the downloading and reading of the data will take time due to the amount of data, a caching method is used. The code checks if the data has been loaded in the local environment and if the data file has been downloaded. Only if the file has not been downloaded or read into R is the relevant operation executed.

```
if (!exists("data.raw")) {

    setwd("./")

    data.path <- "./repdata-data-StormData.csv.bz2"
    if (length(list.files(, pattern = "repdata-data-StormData.csv.bz2",
                          recursive = T)) == 0) {
        data.url <- "http://d396qusza40orc.cloudfront.net/
                repdata%2Fdata%2FStormData.csv.bz2"
        download.file(data.url, data.path)
    }

    data.raw <- data.table(read.csv(bzfile(data.path), stringsAsFactors = F))
}
```

To preserve the original data for backup purposes, a copy of the data is made. Transformations of data are executed on this copy.

```
data.copy <- data.raw
```

Processing of Event Types

As observed, the event descriptions in the `EVTYPE` column are messy and contain many errors that were likely to have been introduced by human error during data input. This column should be cleaned to match the NOAA official 48 event types, as stated in their "NATIONAL WEATHER SERVICE INSTRUCTION 10-1605" document dated 17th August 2007.

The cleaning procedure as used was recommended by David Hood, community TA in the Peer Assessment 2 discussion forums of the Reproducible Research class on Coursera.

Firstly, the text is "squashed", changing all alphabetic characters to lower case and stripping out any spaces as well as punctuation. This process will make it easier for sorting each item into the official 48 event types.

```
data.copy$EVTYPE <- tolower(data.copy$EVTYPE)
data.copy$EVTYPE <- gsub("[[:space:]]*|[[:punct:]]*", "", data.copy$EVTYPE)
```

Unfortunately, there is no encompassing, programmatic solution to replacing the messy event types in the raw data with the official 48 types. Therefore, the user will have to replace the raw data with the official type using a series of regular expression to subset and replace in a separate column. That column is first initialised as seen in the first line.

It should be acknowledged here that the order of the subsetting and the regular expressions used play a big part in the sorting of entries into the pigeonholes of the official event types. The process is heavily dependent on the intuition of the user and could be better fine-tuned with more time and discussion of the individual raw event descriptions.

```
data.copy$cleanev <- ""
data.copy$cleanev[grep("astronomical", data.copy$EVTYPE)] <-
    "Astronomical Low Tide"
data.copy$cleanev[grep("avalan", data.copy$EVTYPE)] <-
    "Avalanche"
data.copy$cleanev[grep("blizz", data.copy$EVTYPE)] <-
    "Blizzard"
data.copy$cleanev[grep("coast", data.copy$EVTYPE)] <-
    "Coastal Flood"
data.copy$cleanev[grep("w(i)?nd", data.copy$EVTYPE)] <-
    "High Wind"
data.copy$cleanev[grep("^cold|^cool", data.copy$EVTYPE)] <-
    "Cold/Wind Chill"
data.copy$cleanev[grep("debris", data.copy$EVTYPE)] <-
    "Debris Flow"
data.copy$cleanev[grep("fog", data.copy$EVTYPE)] <-
    "Dense Fog"
data.copy$cleanev[grep("smoke", data.copy$EVTYPE)] <-
    "Dense Smoke"
data.copy$cleanev[grep("drought", data.copy$EVTYPE)] <-
    "Drought"
data.copy$cleanev[grep("dustd", data.copy$EVTYPE)] <-
    "Dust Devil"
data.copy$cleanev[grep("duststr", data.copy$EVTYPE)] <-
    "Dust Storm"
```

```
data.copy$cleanev[grep("heat", data.copy$EVTYPE)] <-
    "Heat"
data.copy$cleanev[grep("excessiveheat|extremeh", data.copy$EVTYPE)] <-
    "Excessive Heat"
data.copy$cleanev[grep("extremec|extremer|extremew", data.copy$EVTYPE)] <-
    "Extreme Cold/Wind Chill"
data.copy$cleanev[grep("flash", data.copy$EVTYPE)] <-
    "Flash Flood"
data.copy$cleanev[grep("^flood", data.copy$EVTYPE)] <-
    "Flood"
data.copy$cleanev[grep("frost|freez", data.copy$EVTYPE)] <-
    "Frost/Freeze"
data.copy$cleanev[grep("funnel", data.copy$EVTYPE)] <-
    "Funnel Cloud"
data.copy$cleanev[grep("freezingfog", data.copy$EVTYPE)] <-
    "Freezing Fog"
data.copy$cleanev[grep("hail", data.copy$EVTYPE)] <-
    "Hail"
data.copy$cleanev[grep("rain|tstm", data.copy$EVTYPE)] <-
    "Heavy Rain"
data.copy$cleanev[grep("snow", data.copy$EVTYPE)] <-
    "Heavy Snow"
data.copy$cleanev[grep("surf", data.copy$EVTYPE)] <-
    "High Surf"
data.copy$cleanev[grep("hurricane|typhoon", data.copy$EVTYPE)] <-
    "Hurricane (Typhoon)"
data.copy$cleanev[grep("icestorm", data.copy$EVTYPE)] <-
    "Ice Storm"
data.copy$cleanev[grep("lakeeffectsnow", data.copy$EVTYPE)] <-
    "Lake-Effect Snow"
data.copy$cleanev[grep("lake(shore)?flood", data.copy$EVTYPE)] <-
    "Lakeshore Flood"
data.copy$cleanev[grep("lightning", data.copy$EVTYPE)] <-
    "Lightning"
data.copy$cleanev[grep("marinehail", data.copy$EVTYPE)] <-
    "Marine Hail"
data.copy$cleanev[grep("marinehighwind", data.copy$EVTYPE)] <-
    "Marine High Wind"
data.copy$cleanev[grep("marinestormwind", data.copy$EVTYPE)] <-
    "Marine Strong Wind"
data.copy$cleanev[grep("marinethunderstormwind|marinetstmwind",
                       data.copy$EVTYPE)] <- "Marine Thunderstorm Wind"
data.copy$cleanev[grep("ripcurrent", data.copy$EVTYPE)] <-
    "Rip Current"
data.copy$cleanev[grep("seiche", data.copy$EVTYPE)] <-
    "Seiche"
data.copy$cleanev[grep("sleet", data.copy$EVTYPE)] <-
    "Sleet"
data.copy$cleanev[grep("stormsurge", data.copy$EVTYPE)] <-
    "Storm Surge/Tide"
data.copy$cleanev[grep("strongwind", data.copy$EVTYPE)] <-
    "Strong Wind"
data.copy$cleanev[grep("storm(.*)?wind|tstmw", data.copy$EVTYPE)] <-
```

```
    "Thunderstorm Wind"
data.copy$cleanev[grep("tornado", data.copy$EVTYPE)] <-
    "Tornado"
data.copy$cleanev[grep("tropicaldepression", data.copy$EVTYPE)] <-
    "Tropical Depression"
data.copy$cleanev[grep("tropicalstorm", data.copy$EVTYPE)] <-
    "Tropical Storm"
data.copy$cleanev[grep("tsunami", data.copy$EVTYPE)] <-
    "Tsunami"
data.copy$cleanev[grep("volcanic", data.copy$EVTYPE)] <-
    "Volcanic Ash"
data.copy$cleanev[grep("wa.erspout", data.copy$EVTYPE)] <-
    "Waterspout"
data.copy$cleanev[grep("wild(.*)?fire", data.copy$EVTYPE)] <-
    "Wildfire"
data.copy$cleanev[grep("wint", data.copy$EVTYPE)] <-
    "Winter Weather"
data.copy$cleanev[grep("winterst", data.copy$EVTYPE)] <-
    "Winter Storm"
data.copy$cleanev[grep("^$|^[[:space:]]$", data.copy$cleanev)] <- "Other"
```

Processing of Damage Measures

The 2nd major process in the data processing workflow is to enable the proper quantification of economic damage caused by each event. The raw data had listed the amount of damage caused by each event separately based on damage to property or agricultural products in `PROPDMG` and `CROPDMG`, with factors listed in `PROPDMGEXP` and `CROPDMGEXP`.

To obtain the correct amount of damage numerically, the amounts in the numerical columns (`PROPDMG`, `CROPDMG`) should be multiplied against the factors as represented by the character columns (`PROPDMGEXP`, `CROPDMGEXP`). This was as understood from the instructions document, together with representation of the letters. However, since the latter 2 columns contained a mix of alphanumeric characters in mixed case, this needed to be cleaned prior to any further transformation. Here, all characters were coerced into upper case and digits as well as punctuation was stripped out.

```
data.copy$PROPDMGEXP <- toupper(data.copy$PROPDMGEXP)
data.copy$CROPDMGEXP <- toupper(data.copy$CROPDMGEXP)
data.copy$PROPDMGEXP <- gsub("[[:digit:]]|[[:punct:]]", "",
                             data.copy$PROPDMGEXP)
data.copy$CROPDMGEXP <- gsub("[[:digit:]]|[[:punct:]]", "",
                             data.copy$CROPDMGEXP)
```

Once the "exponent" columns have been cleaned, the letters are substituted with the factors they represent for easier multiplication after they are coerced to numeric type. Here, it is assumed that "H" represents a factor of 100, and blanks represent a factor of 1.

```
data.copy$PROPDMGEXP <- gsub("^$", 1, data.copy$PROPDMGEXP)
data.copy$PROPDMGEXP <- gsub("H", 100, data.copy$PROPDMGEXP)
data.copy$PROPDMGEXP <- gsub("K", 1000, data.copy$PROPDMGEXP)
data.copy$PROPDMGEXP <- gsub("M", 1000000, data.copy$PROPDMGEXP)
data.copy$PROPDMGEXP <- gsub("B", 1000000000, data.copy$PROPDMGEXP)
```

```
data.copy$CROPDMGEXP <- gsub("^$", 1, data.copy$CROPDMGEXP)
data.copy$CROPDMGEXP <- gsub("H", 100, data.copy$CROPDMGEXP)
data.copy$CROPDMGEXP <- gsub("K", 1000, data.copy$CROPDMGEXP)
data.copy$CROPDMGEXP <- gsub("M", 1000000, data.copy$CROPDMGEXP)
data.copy$CROPDMGEXP <- gsub("B", 1000000000, data.copy$CROPDMGEXP)

data.copy$PROPDMGEXP <- as.numeric(data.copy$PROPDMGEXP)
data.copy$CROPDMGEXP <- as.numeric(data.copy$CROPDMGEXP)
```

Ranking and Summary

Once the data has been suitably cleaned, it can be summarised to obtain averaged measures of the deaths/injuries and economic damage by event type. The data is first mutated to produce fully quantitative measures of economic damage, then is summarised by averaging each measure by event type. Lastly a final mutation is done to obtain the totals of each type of measure (human health hazards, economic damage), for use later in ranking.

```
data.damage <- data.copy %>%
                mutate(pdmg = PROPDMG * PROPDMGEXP,
                       cdmg = CROPDMG * CROPDMGEXP,
                       human = FATALITIES + INJURIES,
                       econ = pdmg + cdmg) %>%
                group_by(cleanev) %>%
                summarise(human = round(mean(human)),
                          econ = mean(econ))
```

As there are 48 types of events, it is likely that the plots will be very cluttered with such highly dimensional data. Therefore, the following code seeks to only identify the top 10 event types based on the hazards to human health and economic damage respectively.

```
data.human <- data.damage[order(data.damage$human, decreasing=T)][1:10]

data.econ <- data.damage[order(data.damage$econ, decreasing=T)][1:10]
```

Results

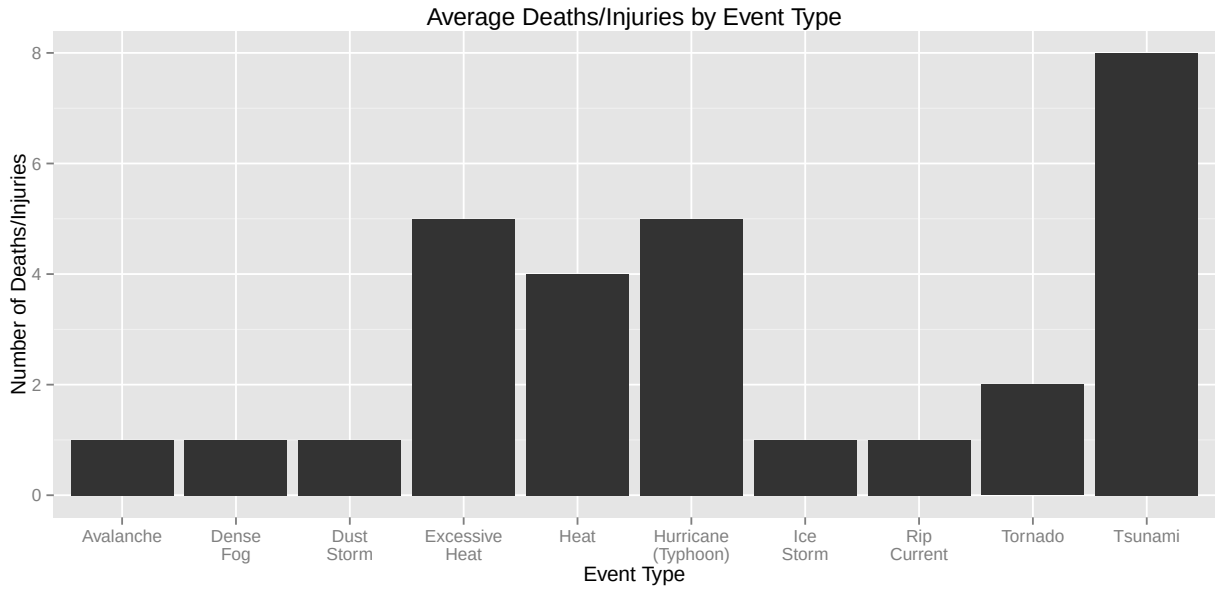
To display the results neatly, the `ggplot2` and `scales` packages are required.

```
if (!suppressMessages(require("ggplot2"))){
    install.packages("ggplot2")
    suppressMessages(require("ggplot2"))
}
if (!suppressMessages(require("scales"))){
    install.packages("scales")
    suppressMessages(require("scales"))
}
```

Events Most Harmful to Human Health

The following code creates a plot of the top 10 events most harmful to human health with bars representing the average deaths/injuries per event.

```
graph.human <- ggplot(data.human, aes(x = gsub("[[:space:]]", "\n", cleanev),
                                      y = human)) +
               geom_bar(stat = "identity") +
               labs(title = "Average Deaths/Injuries by Event Type",
                    x = "Event Type",
                    y = "Number of Deaths/Injuries") +
               scale_y_continuous(labels = comma)
print(graph.human)
```

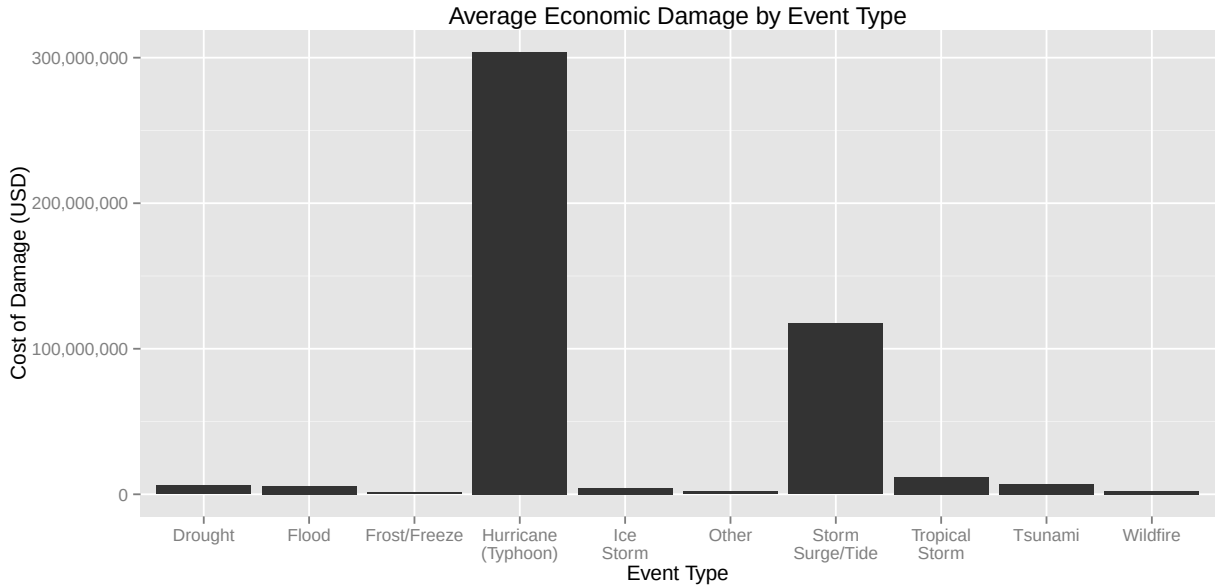


As seen in the plot above, the events that pose the greatest threat to human health are Tsunami, with an average of 8 deaths/injuries per event.

Events with Greatest Economic Consequences

The following code creates a plot of the top 10 events causing the most economic damage with bars representing the average economic damage per event.

```
graph.econ <- ggplot(data.econ, aes(x = gsub("[[:space:]]", "\n", cleanev),
                                    y = econ)) +
              geom_bar(stat = "identity") +
              labs(title = "Average Economic Damage by Event Type",
                   x = "Event Type",
                   y = "Cost of Damage (USD)") +
              scale_y_continuous(labels = comma)
print(graph.econ)
```



As seen in the plot above, the events that cause the greatest economic damage are Hurricane (Typhoon), with an average of approximately 303,921,498 USD worth of damage per event.

Conclusion

In terms of the threat to human health, it can be seen that most events to do not pose much a threat, given an average of only 1 - 2 deaths/injuries per event. However, the economic consequences can vary greatly from event to event. It is not clear if the the economic consequences for other events are as significant as Hurricane (Typhoon) from the plot, however, it can be studied in greater detail with the full averaged data in `data.damage` included in the appendix.

Appendix - data.damage data

```
if (!suppressMessages(require("xtable"))){
    install.packages("xtable")
    suppressMessages(require("xtable"))
}
```

```
print(xtable(data.damage), comment = F, include.rownames = F)
```

cleanev	human	econ
Tornado	2.00	971548.52
Thunderstorm Wind	0.00	32732.08
Hail	0.00	65854.35
Heavy Rain	0.00	341789.16
Heavy Snow	0.00	66111.42
Ice Storm	1.00	4413209.33
Winter Storm	0.00	592818.92
Hurricane (Typhoon)	5.00	303921497.69
Other	0.00	1968827.71
Lightning	0.00	60311.74
Dense Fog	1.00	12427.60
Rip Current	1.00	209.78
Flash Flood	0.00	330141.91
High Wind	0.00	299191.16
Funnel Cloud	0.00	27.86
Heat	4.00	450610.69
Flood	0.00	5777338.59
Cold/Wind Chill	0.00	115130.43
Waterspout	0.00	15725.19
Extreme Cold/Wind Chill	0.00	742201.16
Blizzard	0.00	283280.92
Frost/Freeze	0.00	1310205.33
Coastal Flood	0.00	496456.09
Avalanche	1.00	9617.05
Excessive Heat	5.00	294278.67
Dust Storm	1.00	20160.84
Sleet	0.00	16393.44
Dust Devil	0.00	4790.87
Strong Wind	0.00	66190.82
High Surf	0.00	109619.00
Wildfire	0.00	2102185.09
Winter Weather	0.00	5124.20
Drought	0.00	6036443.73
Storm Surge/Tide	0.00	117275254.28
Tropical Storm	1.00	12064973.53
Lakeshore Flood	0.00	315416.67
Lake-Effect Snow	0.00	60974.20
Freezing Fog	0.00	47434.78
Volcanic Ash	0.00	17241.38
Seiche	0.00	46666.67
Tropical Depression	0.00	28950.00
Dense Smoke	0.00	4761.90
Astronomical Low Tide	0.00	35180.51
Marine Hail	0.00	9.05
Marine High Wind	0.00	9607.48
Tsunami	8.00	7204100.00