Reproducible Research: Peer Assessment 2

Owusu

March 4, 2016

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Analysis of Impact Different Weather Events on Economics and Public Health

1. Assignment

The basic goal of this assignment is to explore the NOAA Storm Database and answer some basic questions about severe weather events. You must use the database to answer the questions below and show the code for your entire analysis. Your analysis can consist of tables, figures, or other summaries. You may use any R package you want to support your analysis.

2. Synopsis

The National Oceanic and Atmospheric Administration (NOAA) maintains a public database for storm event. The data contains the type of storm event, details like location, date, estimates for damage to property as well as the number of human victims of the storm. In this report we investigate which type of events are the most harmful to the population and financially.

The conclusion is that the impact on humans, be it injuries or fatalities, isn't directly correlated to the ecomomic damage weather events cause.

3. Data Processing

```
setwd("/home/owusu/Desktop/Assiignment2")
```

3.1. Load Libraries

Necessary libraries to perform loading, computation, transformation and plotting of data.

```
library(data.table) #for data reading

library(plyr) # for count & aggregate method
library(dplyr) # for select
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from 'package:data.table':
##
##     between, last

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2) # for plots
library(grid) # for grids
library(gridExtra) # for advanced plots
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

3.2. Load Data

Read the source.csv file

```
CsvStormData <- read.csv(bzfile("repdata_data_StormData.csv.bz2"))
```

3.3. Managing the Outliers

An analysis of the biggest outliers follows, to avoid that a small typo might change completely the meaning of the database.

```
# Columns 26 and 28 contain letters that indicate if the money in
# Columns 25 and 27 are hundreds, thousands, millions or billions.
# A control on the billions is needed to avoid big impacts on the
# overall results.

# Property and Crop damages (billions of dollars)
BillionsPIndex <- CsvStormData[,26] == "B"
BillionsCIndex <- CsvStormData[,28] == "B"
BillionsP <-select(CsvStormData[BillionsPIndex,], c(2,7,8,23,24,25,26,27,28))
BillionsC <-select(CsvStormData[BillionsCIndex,], c(2,7,8,23,24,25,26,27,28))
# Check top 10 expenses
TopBilProp <- top_n(BillionsP,10,PROPDMG)
TopBilCrop <- top_n(BillionsC,10,PROPDMG)
TopBilProp
```

```
##              BGN_DATE STATE            EVTYPE FATALITIES INJURIES PROPDMG
## 1   3/12/1993 0:00:00    AL      WINTER STORM          4        0    5.00
## 2   8/31/1993 0:00:00    IL       RIVER FLOOD          0        0    5.00
## 3    6/5/2001 0:00:00    TX    TROPICAL STORM         22        0    5.15
## 4   8/13/2004 0:00:00    FL HURRICANE/TYPHOON          7      780    5.42
## 5  10/24/2005 0:00:00    FL HURRICANE/TYPHOON          5        0   10.00
## 6   8/28/2005 0:00:00    LA HURRICANE/TYPHOON          0        0   16.93
## 7   8/29/2005 0:00:00    LA       STORM SURGE          0        0   31.30
## 8   8/28/2005 0:00:00    MS HURRICANE/TYPHOON          0        0    7.35
## 9   8/29/2005 0:00:00    MS       STORM SURGE          0        0   11.26
## 10  8/29/2005 0:00:00    MS HURRICANE/TYPHOON         15      104    5.88
## 11   1/1/2006 0:00:00    CA             FLOOD          0        0  115.00
##    PROPDMGEXP CROPDMG CROPDMGEXP
## 1           B    0.00
## 2           B    5.00          B
## 3           B    0.00
## 4           B  285.00          M
## 5           B    0.00
## 6           B    0.00
## 7           B    0.00
## 8           B    0.00
## 9           B    0.00
## 10          B    1.51          B
## 11          B   32.50          M
```

```
TopBilCrop
```

```
##            BGN_DATE STATE            EVTYPE FATALITIES INJURIES PROPDMG
## 1 8/20/1995 0:00:00    AL              HEAT          0        0    0.00
## 2 8/31/1993 0:00:00    IL       RIVER FLOOD          0        0    5.00
## 3  8/1/1995 0:00:00    IA           DROUGHT          0        0    0.00
## 4 9/21/1995 0:00:00    IA            FREEZE          0        0    0.00
## 5  2/9/1994 0:00:00    MS         ICE STORM          0        0  500.00
## 6 8/29/2005 0:00:00    MS HURRICANE/TYPHOON         15      104    5.88
## 7  1/1/2006 0:00:00    TX           DROUGHT          0        0    0.00
## 8  9/1/2011 0:00:00    GA           DROUGHT          0        0    0.00
## 9  9/1/2011 0:00:00    GA           DROUGHT          0        0    0.00
##   PROPDMGEXP CROPDMG CROPDMGEXP
## 1               0.40          B
## 2          B    5.00          B
## 3               0.50          B
## 4               0.20          B
## 5          K    5.00          B
## 6          B    1.51          B
## 7               1.00          B
## 8          K    0.00          B
## 9          K    0.00          B
```

```
# Checking on the 1/1/2006 FLOOD that is supposed to cost 115 B in
# Damage properties, it's clear that it's a typo being the estimates
# Around 300 millions. Source: 1 - 300 millions - http://pubs.usgs.gov/of/2006/1182/pdf/ofr2006-1182.pd
# Therefore correct B to M
CsvStormData[605953,26] <- "M"
```

3.4. Transforming the Data

All symbols in the DMGEXP columns are treated as powers of 10 of the DMG column. A calculation of the Property and Crop damages is needed before further analysis, because it allows a proper subsetting of the raw data.

```
StormData <- CsvStormData
PEXP <- revalue(StormData$PROPDMGEXP, c("-"=NA, "?"=NA, "+"=NA,
                                        "h"=2, "H"=2, "K"=3, "m"=6,
                                        "M"=6, "B"=9))
CEXP <- revalue(StormData$CROPDMGEXP, c("?"=NA, "K"=3, "k"=3,
                                        "m"=6, "M"=6, "B"=9))
PEXP <- as.numeric(levels(PEXP)[PEXP])
CEXP <- as.numeric(levels(CEXP)[CEXP])
StormData$PROPDMG <- StormData$PROPDMG * (10^PEXP)
StormData$CROPDMG <- StormData$CROPDMG * (10^CEXP)
```

3.5. Subsetting the Dataset

In our analysis we are interested in few of the 37 variables of the dataset: fatalities, injuries, damage on properties and damage on crops. Subsetting the dataset when one or more of these are positive, makes the database smaller and the analysis easier.

```
StormData <- subset(StormData, FATALITIES>0 | INJURIES>0 | PROPDMG>0 | CROPDMG>0)
```

3.6.Refactor EVTYPE

EVTYPES variable contains the name of the events. This variable in its raw form is very confusing and contains many typos. An elaboration on the variable is needed to remove punctuation and extra spaces from the data.

```
# recalculate factor levels after subsetting
Types <- levels(StormData$EVTYPE)[StormData$EVTYPE]
OldTypes <- Types <- levels(as.factor(Types))
# all uppercase
Types <- toupper(Types)
# replace punctuation with spaces
Types <- gsub('([[:punct:]])|\\s+',' ',Types)
# remove leading and trailing spaces
Types <- gsub("^\\s+|\\s+$", "", Types)
# replace double spaces with one space
Types <- gsub("  ", " ",Types)
```

Considering the clean EVTYPE variable, many events have different names that fall in analogue categories.

3.7. Clustering Process

```
# First clustering process
Clusters <- c("FLOOD","COLD","HAIL","SNOW","HURRICANE","ICE",
              "LIGHTNING", "MARINE", "THUNDERSTORM", "TSTM", "TORNADO",
```

```
              "FIRE", "MUD", "WIND", "STORM", "DUST", "FREEZ",
              "HEAT", "RAIN", "SURF", "SURGE", "FOG", "BLIZZARD",
              "HYPOTHERMIA", "SLIDE", "TIDE", "EROSION", "WARM",
              "PRECIP", "SWELL", "SEA", "LOWTEMP", "WINTER",
              "WATER", "WINTRY", "WAVE", "SHOWER", "SLUMP",
              "HYPERTHERMIA", "GLAZE", "URBAN", "FROST", "COOL",
              "WET", "BURST", "SMOKE", "TURBULENCE")
for (i in 1:length(Clusters)){
    Types[grep(Clusters[i],Types)] <- Clusters[i]
}

# Typos
Types[Types == "LIGHTING" | Types == "LIGNTNING"] <- "LIGHTNING"
Types[Types == "AVALANCE"] <- "AVALANCHE"
Types[Types == "TORNDAO"] <- "TORNADO"

# Second granular clustering process
Types[Types == "TYPHOON"] <- "HURRICANE"
Types[Types == "WATER"] <- "FLOOD"
Types[Types == "SLEET"] <- "SNOW"
Types[Types == "GLAZE" | Types == "ICY ROADS" |
          Types == "FROST"] <- "ICE"
Types[Types == "SLIDE" | Types == "LANDSPOUT"] <- "SLUMP"
Types[Types == "PRECIP" | Types == "SHOWER" | Types == "WET"] <- "RAIN"
Types[Types == "WARM" | Types == "HYPERTHERMIA"] <- "HEAT"
Types[Types == "FREEZ" | Types == "HYPOTHERMIA" |
          Types == "WINTER" | Types == "WINTRY" |
          Types == "LOW TEMPERATURE" | Types == "COOL"] <- "COLD"
Types[Types == "TSTM" | Types == "THUNDERSTORM" |
          Types == "TURBULENCE"] <- "STORM"
Types[Types == "SURF" | Types == "SURGE" | Types == "TSUNAMI" |
      Types == "SWELL" | Types == "SEA" | Types =="WAVE" |
          Types == "SEICHE"] <- "WAVES"
Types[Types == "" | Types == "APACHE COUNTY" | Types == "DAM BREAK" |
          Types=="HEAVY MIX" | Types=="RIP CURRENT" |
          Types=="RIP CURRENTS" | Types=="HIGH"] <- "OTHER"
```

3.8. Creating Dataset and variables for plottings

Replace levels in StormData using the new clusters.

```
StormData$EVTYPE <- droplevels(StormData$EVTYPE,OldTypes)
StormData$EVTYPE <- mapvalues(StormData$EVTYPE, from = OldTypes, to = Types)
```

Calculate Fatalities and Injuries provoked by the different events.

```
Types <- levels(StormData$EVTYPE)
Fatalities <- tapply(StormData$FATALITIES, StormData$EVTYPE, sum)
Injuries <- tapply(StormData$INJURIES, StormData$EVTYPE, sum)
HarmData <- as.data.table(cbind(Types,Fatalities,Injuries))
HarmData$Fatalities <- as.integer(HarmData$Fatalities)
HarmData$Injuries <- as.integer(HarmData$Injuries)
HarmData <- HarmData[HarmData$Fatalities > 0 | HarmData$Injuries > 0,]
```

Calculate properties & crop damage provoked by the different events.

```
# Transform NA into 0 to allow function SUM to work properly
StormData$PROPDMG[is.na(StormData$PROPDMG)] <- 0
StormData$CROPDMG[is.na(StormData$CROPDMG)] <- 0
# Calculate damages per event
Pdamage <- tapply(StormData$PROPDMG, StormData$EVTYPE, sum)
Cdamage <- tapply(StormData$CROPDMG, StormData$EVTYPE, sum)
TotalCost <- Pdamage + Cdamage
PD <- cbind(type=Types,amount=Pdamage,damage="Property Damage",
            total=TotalCost)
CD <- cbind(type=Types,amount=Cdamage,damage="Crop Damage",
            total=TotalCost)
TotalCost <- as.data.frame(TotalCost[order(-TotalCost)])
TotalCost$lvs <- rownames(TotalCost)
Top10TC <- head(TotalCost,10)[,2]
Damage <- as.data.table(rbind(PD,CD))
Damage$amount <- as.numeric(Damage$amount)
Damage$total <- as.numeric(Damage$total)
Damage <- top_n(Damage,20,total)
Damage$type <- factor(Damage$type, levels = Top10TC)
```

4. Results

The results of the Data Analysis address the following questions: * 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?

4.1 The most harmful with respect to population health across the United States.

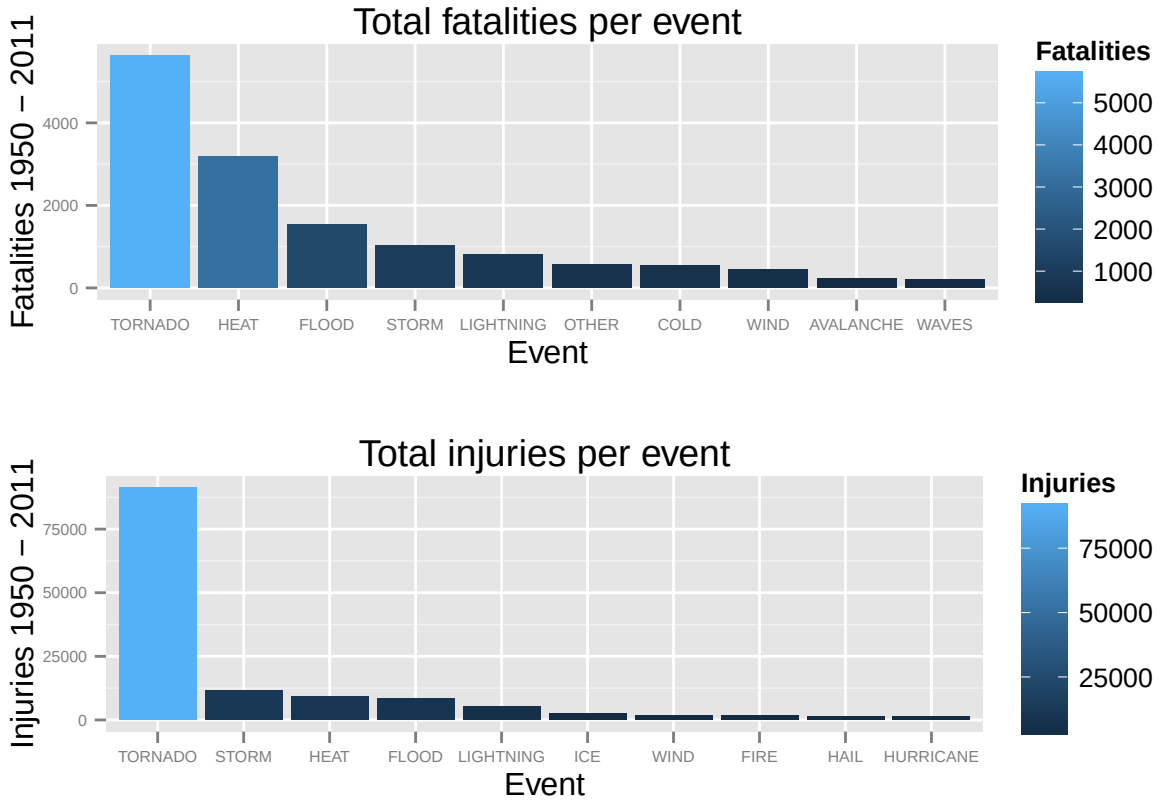
To show how harmful is a specific cluster of events on the US population, two different charts are displayed, that is fatalities and injuries. These two statistics are different in terms of quality and can't be aggregated. For this reason it is important to observe them separately.

Fatalities and Injuries Plot

```
# Select top 10 fatalities and injuries
FatData <- top_n(HarmData,10,Fatalities)
FatData <- FatData[order(-Fatalities)]
FatData$Types <- factor(FatData$Types, levels = FatData$Types)
InjData <- top_n(HarmData,10,Injuries)
InjData <- InjData[order(-Injuries)]
InjData$Types <- factor(InjData$Types, levels = InjData$Types)

FatChart <- ggplot(FatData, aes(x=Types, y=Fatalities, fill=Fatalities)) + geom_bar(stat = "identity") +
    ylab("Fatalities 1950 - 2011") +
    xlab("Event") +
    ggtitle ("Total fatalities per event") +
    theme(axis.text=element_text(size=6))
InjChart <- ggplot(InjData, aes(x=Types, y=Injuries, fill=Injuries)) + geom_bar(stat = "identity") +
    ylab("Injuries 1950 - 2011") +
    xlab("Event") +
    ggtitle ("Total injuries per event") +
```

```
    theme(axis.text=element_text(size=6))
grid.arrange(FatChart, InjChart, nrow=2)
```

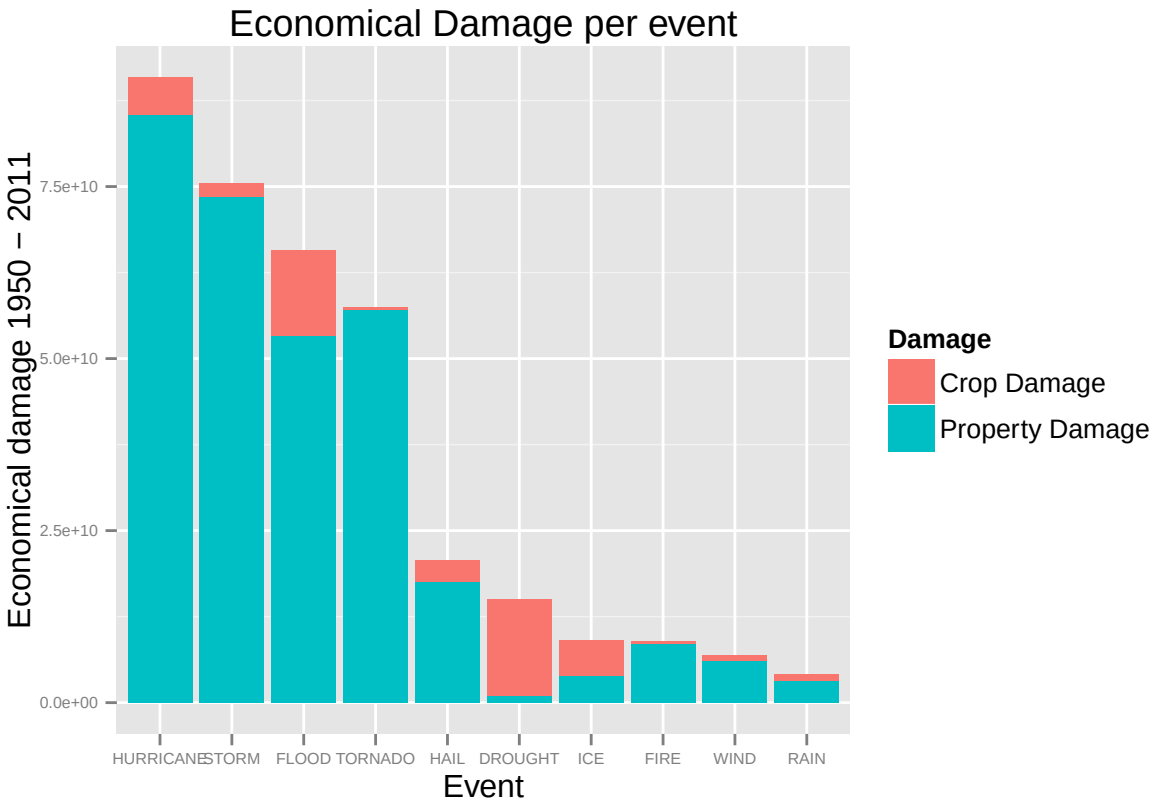


It can be observed that tornados are the first event that determines both fatalities and injuries. The top 5 of the most harmful events in both fatalities and injuries are: Tornados, Heat, Floods, Storms and Lightnings. Considering that Hurricanes are one of the most frequent carriers of Tornados, Floods, Storms and Lightnings, it would be reasonable to think that actually hurricanes are the first cause of harm in the US. Heat is the only event not strichtly related to hurricanes in the top 5.

4.2 Events that have the greatest economic consequences across the United States.

Economical damage is a quantitative variable measured in dollars. The database has two kinds of economical damage: the proprerty damage and the crop damage. Being both quantiative, this analysis calculates their sum showing in the plot how this total is distributed among properties and crops.

```
EconomicalChart <- ggplot(Damage, aes(x=type, y=amount, fill=factor(damage))) +
    geom_bar(stat = "identity") +
    ylab("Economical damage 1950 - 2011") +
    xlab("Event") +
    scale_fill_discrete(name = "Damage") +
    ggtitle ("Economical Damage per event") +
    theme(axis.text=element_text(size=6))
EconomicalChart
```



Hurricanes, Storms, Floods, Tornados and Hail are the events that bring the biggest economical damage. As considered in the previous plot, storms, tornados and floods are many times part of the hurricanes. For this reason, we can consider Hurricanes the biggest threat for United States economy, like Katrina demonstrated in 2005. It's worth noticing that the #1 factor for Crop Damage is actually Drought, an event that shouldn't be underesitmated especially in the warmest countries of the United States.

5. Conclusions

- Hurricanes, Tornados, Storms and Floods are the key events that threaten the safety and economics of the United States. The Government should invest wisely to prevent and protect the citizens from these events.
- Drought are the biggest threat for the agricultural economy, and water plans should be designed in every state at risk.