# Stormdata_assg2.Rmd

*Ranu*

*July 29, 2016*

### Introduction

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, and property damage, and preventing such outcomes to the extent possible is a key concern.

This project involves exploring the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database. This database tracks characteristics of major storms and weather events in the United States, including when and where they occur, as well as estimates of any fatalities, injuries, and property damage.

### 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. These data show fatalities and injuries for each event as well as the economic damage caused on crops and properties. The total effect of each type of event over the years is considered for analysis.In this report we investigate which type of events are the most harmful to the population both health-wise and econimicially. The results show that *TORNADO* is the event that causes most harmful effects to population health.The event *FLOOD* and *DROUGHT* cause most economic consequences i.e, propert damage and crop damage respectively.

### Data Processing

The data for this assignment come in the form of a comma-separated-value file compressed via the bzip2 algorithm to reduce its size. The data was downloaded from the Coursera Reproducible Research web site : Stormdata (47Mb) This dataset consists of 902297 obs. of 37 variables. We do not need all thevariables for our analysis purpose. The variables that are required are: EVTYPE ,FATALITIES, INJURIES, PROPDMG, PROPDMGEXP, CROPDMG, CROPDMGEXP. We shall be subseting the following comlumns from the dataset. Event type (EVTYPE) entries that are similar have been grouped together. The events in the database start in the year 1950 and end in November 2011.

```
#downloading the compressed file
assgdata<- "repdata_data_StormData.csv.bz2"
if (!file.exists(assgdata)) {
    fileURL<- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
    download.file(fileURL,assgdata,mode = "wb")
}
# load data into R
stormData <- read.csv("repdata_data_StormData.csv.bz2")
head(stormData)
```

```
##   STATE__           BGN_DATE BGN_TIME TIME_ZONE COUNTY COUNTYNAME STATE
## 1       1  4/18/1950 0:00:00     0130       CST     97     MOBILE    AL
## 2       1  4/18/1950 0:00:00     0145       CST      3    BALDWIN    AL
```

```
## 3       1  2/20/1951 0:00:00     1600       CST     57    FAYETTE    AL
## 4       1   6/8/1951 0:00:00     0900       CST     89    MADISON    AL
## 5       1 11/15/1951 0:00:00     1500       CST     43    CULLMAN    AL
## 6       1 11/15/1951 0:00:00     2000       CST     77 LAUDERDALE    AL
##    EVTYPE BGN_RANGE BGN_AZI BGN_LOCATI END_DATE END_TIME COUNTY_END
## 1 TORNADO         0                                               0
## 2 TORNADO         0                                               0
## 3 TORNADO         0                                               0
## 4 TORNADO         0                                               0
## 5 TORNADO         0                                               0
## 6 TORNADO         0                                               0
##   COUNTYENDN END_RANGE END_AZI END_LOCATI LENGTH WIDTH F MAG FATALITIES
## 1         NA         0                      14.0   100 3   0          0
## 2         NA         0                       2.0   150 2   0          0
## 3         NA         0                       0.1   123 2   0          0
## 4         NA         0                       0.0   100 2   0          0
## 5         NA         0                       0.0   150 2   0          0
## 6         NA         0                       1.5   177 2   0          0
##   INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP WFO STATEOFFIC ZONENAMES
## 1       15    25.0          K       0
## 2        0     2.5          K       0
## 3        2    25.0          K       0
## 4        2     2.5          K       0
## 5        2     2.5          K       0
## 6        6     2.5          K       0
##   LATITUDE LONGITUDE LATITUDE_E LONGITUDE_ REMARKS REFNUM
## 1     3040      8812       3051       8806              1
## 2     3042      8755          0          0              2
## 3     3340      8742          0          0              3
## 4     3458      8626          0          0              4
## 5     3412      8642          0          0              5
## 6     3450      8748          0          0              6
```

```r
data <- stormData[, c(8,23,24,25,26,27,28)]
str(data)
```

```
## 'data.frame':    902297 obs. of  7 variables:
##  $ EVTYPE    : Factor w/ 985 levels "   HIGH SURF ADVISORY",..: 834 834 834 834 834 834 834 834 834 8
##  $ 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 ...
```

### Extracting of distinct levels of column PROPDMGEXP

```r
unique(data$PROPDMGEXP)
```

```
##  [1] K M   B m + 0 5 6 ? 4 2 3 h 7 H - 1 8
## Levels:  - ? + 0 1 2 3 4 5 6 7 8 B h H K m M
```

**Assigning numeric values to these distinct levels**

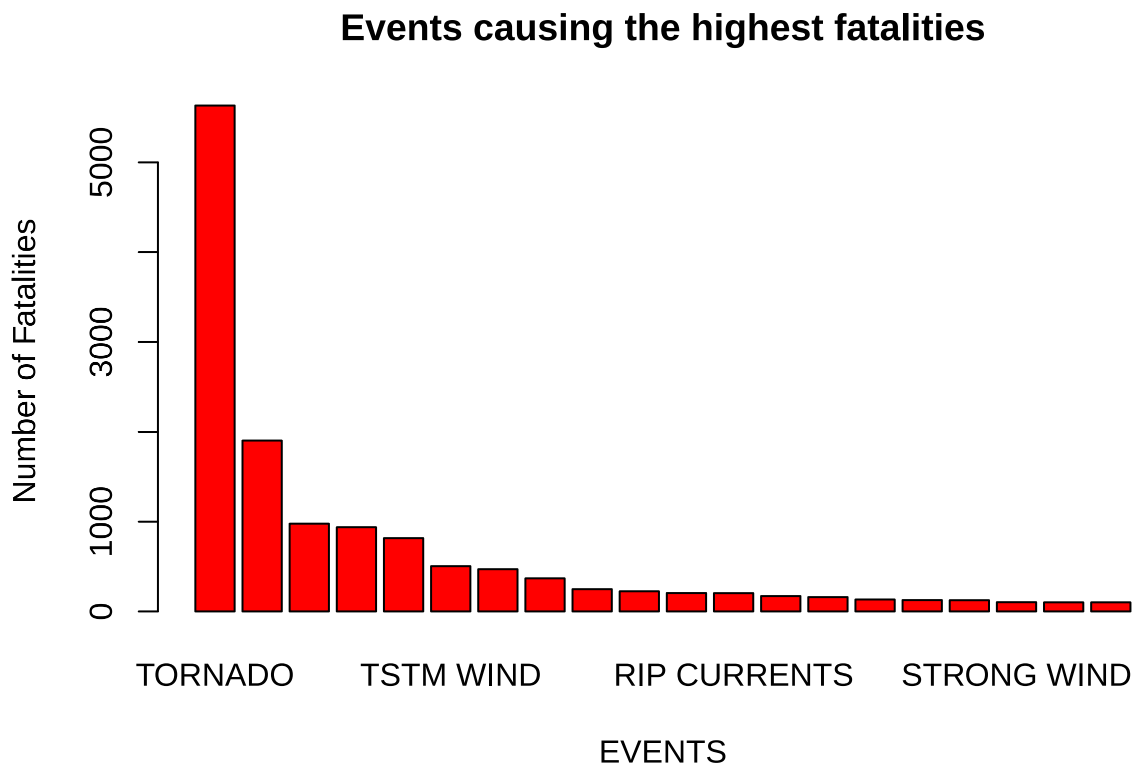
```
data$PROPEXP[data$PROPDMGEXP == "K"] <- 1000
data$PROPEXP[data$PROPDMGEXP == "M"] <- 1000000
data$PROPEXP[data$PROPDMGEXP == ""] <- 1
data$PROPEXP[data$PROPDMGEXP == "B"] <- 1000000000
data$PROPEXP[data$PROPDMGEXP == "m"] <- 1000000
data$PROPEXP[data$PROPDMGEXP == "0"] <- 1
data$PROPEXP[data$PROPDMGEXP == "5"] <- 100000
data$PROPEXP[data$PROPDMGEXP == "6"] <- 1000000
data$PROPEXP[data$PROPDMGEXP == "4"] <- 10000
data$PROPEXP[data$PROPDMGEXP == "2"] <- 100
data$PROPEXP[data$PROPDMGEXP == "3"] <- 1000
data$PROPEXP[data$PROPDMGEXP == "h"] <- 100
data$PROPEXP[data$PROPDMGEXP == "7"] <- 10000000
data$PROPEXP[data$PROPDMGEXP == "H"] <- 100
data$PROPEXP[data$PROPDMGEXP == "1"] <- 10
data$PROPEXP[data$PROPDMGEXP == "8"] <- 100000000
# give 0 to invalid exponent data, so they not count in
data$PROPEXP[data$PROPDMGEXP == "+"] <- 0
data$PROPEXP[data$PROPDMGEXP == "-"] <- 0
data$PROPEXP[data$PROPDMGEXP == "?"] <- 0
# compute the property damage value
data$PROPDMG_TOT <- data$PROPDMG * data$PROPEXP
```

**Assigning numeric values to the levels of CROPDMGEXP**

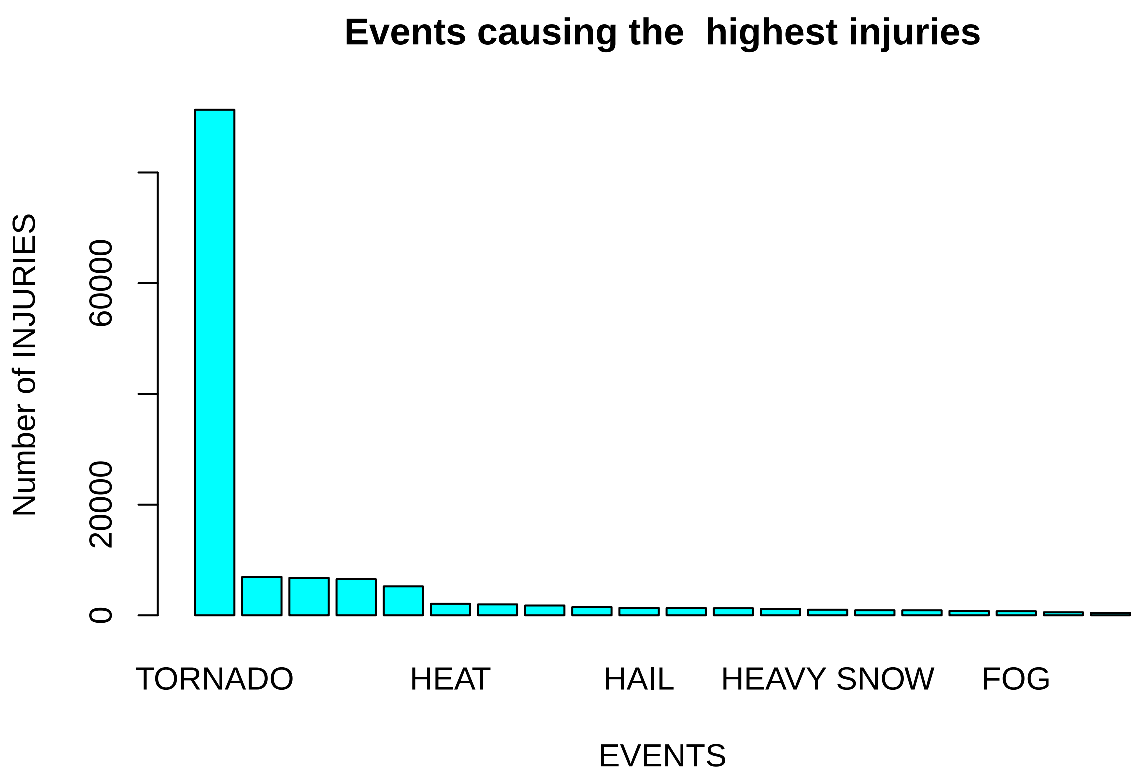
```
data$CROPEXP[data$CROPDMGEXP == "M"] <- 1000000
data$CROPEXP[data$CROPDMGEXP == "K"] <- 1000
data$CROPEXP[data$CROPDMGEXP == "m"] <- 1000000
data$CROPEXP[data$CROPDMGEXP == "B"] <- 1000000000
data$CROPEXP[data$CROPDMGEXP == "0"] <- 1
data$CROPEXP[data$CROPDMGEXP == "k"] <- 1000
data$CROPEXP[data$CROPDMGEXP == "2"] <- 100
data$CROPEXP[data$CROPDMGEXP == ""] <- 1
# give 0 to invalid exponent data, so they not count in
data$CROPEXP[data$CROPDMGEXP == "?"] <- 0
# compute the crop damage value
data$CROPDMG_TOT <- data$CROPDMG * data$CROPEXP
```

**Across the United States, which types of events are most harmful with respect to population health?**

```
#calculating total fatalities based on the event type
tot_fatalities <- aggregate(FATALITIES ~ EVTYPE, data, sum)
fatalities <- tot_fatalities[order(-tot_fatalities$FATALITIES),][1:20,]
#plotting the graph of events ~ no. of fatalities
barplot(fatalities$FATALITIES, names.arg = fatalities$EVTYPE,
        main = "Events causing the highest fatalities",
        ylab = "Number of Fatalities", xlab = "EVENTS", col = "red")
```




```r
#calculating total fatalities based on the event type
tot_injuries <- aggregate(INJURIES ~ EVTYPE, data, sum)
injuries <- tot_injuries[order(-tot_injuries$INJURIES),][1:20,]
#plotting the graph of events ~ no. of injuries
barplot(injuries$INJURIES, names.arg = injuries$EVTYPE,
        main = "Events causing the  highest injuries",
        ylab = "Number of INJURIES", xlab = "EVENTS", col = "cyan")
```



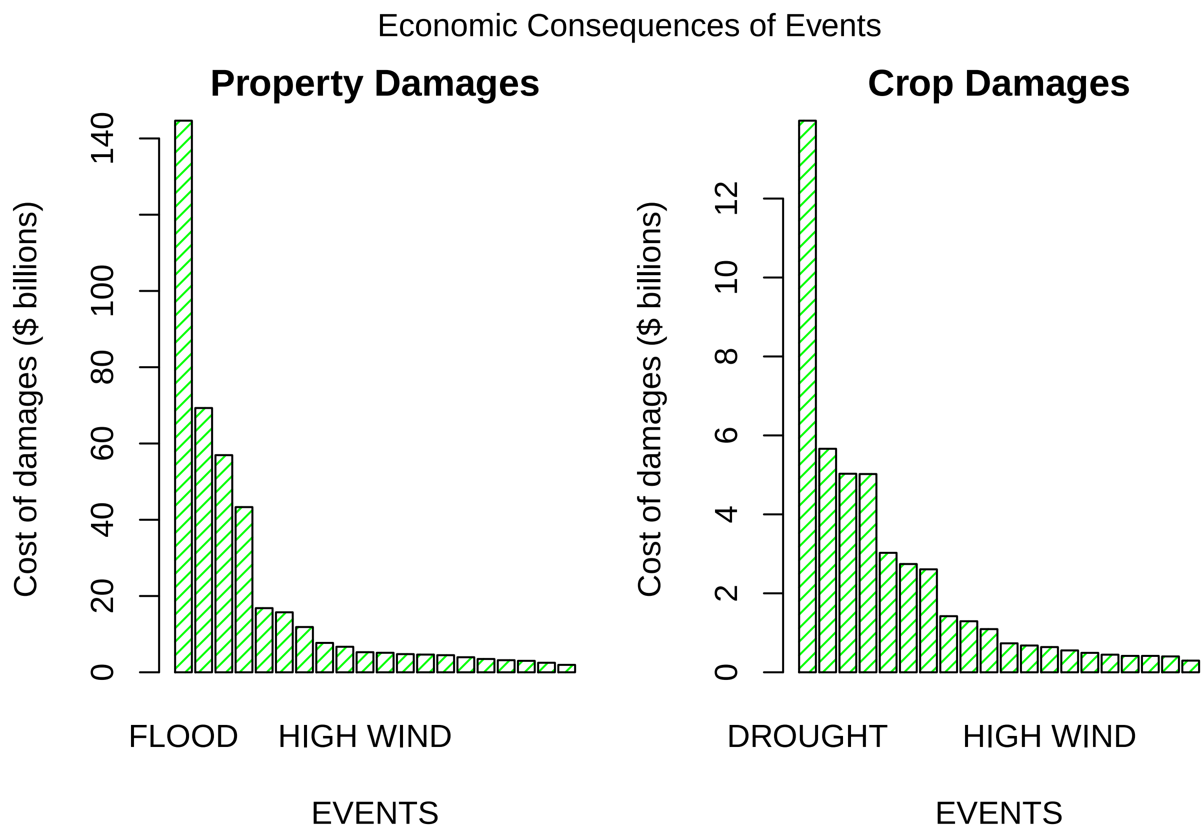

**RESULT :**

From the above plot it can be clealy stated that the event *TORNADO* causes the greatest threat to population health.

**Across the United States, which types of events have the greatest economic consequences?**

```
#calculating total property damage based on the event type
tot_propdmg <- aggregate(PROPDMG_TOT ~ EVTYPE, data, sum)
propdmg <- tot_propdmg[order(-tot_propdmg$PROPDMG_TOT), ][1:20, ]
#calculating total cropdamage based on the event type
tot_cropgmd <- aggregate(CROPDMG_TOT ~ EVTYPE, data, sum)
cropdmg <- tot_cropgmd[order(-tot_cropgmd$CROPDMG_TOT), ][1:20, ]
#making a panel plot of events ~ damages
par(mfrow=c(1,2),mar=c(4,4,2,1),oma=c(0,0,2,0))
barplot(propdmg$PROPDMG_TOT/(10^9), names.arg = propdmg$EVTYPE,
        main = "Property Damages", ylab = "Cost of damages ($ billions)", xlab = "EVENTS", density = 20
barplot(cropdmg$CROPDMG_TOT/(10^9), names.arg = cropdmg$EVTYPE,
    main = "Crop Damages", ylab = "Cost of damages ($ billions)", xlab = "EVENTS", density = 20, col =

mtext("Economic Consequences of Events", outer = TRUE)
```




**RESULT :**

From the above plot it can be clealy stated that the event *FLOOD* causes greatest property damage and the event *DROUGHT* causes greatest crop damage.