Reproducible Research: Peer Assessment 2

Most Harmful Weather Events

The purpose of the analysis is to determine which types of events are most harmful with respect to population health in the United States by using the NOAA Storm Database. The paper attempts to answer two basic questions about severe weather events:

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?

Synopsis

The analysis finds that...

Data Processing

I begin the analysis by loading libraries and setting a few global parameters:

```
## load needed libraries, set global options, and working directory
library(knitr); library(plyr)
opts_chunk$set(echo=TRUE)
setwd("~/Documents/Courses/datasciencecoursera/RepResProj2/")
```

We first download and unzip the data (if necessary):

```
#Download file if it does not exist
if (!file.exists("repdata-data-StormData.csv.bz2")) {
    fileURL <- "http://bit.ly/1uNSAQY"
    zipfile = "repdata-data-StormData.csv.bz2"
    download.file(fileURL, destfile=zipfile, method="curl")
}
```

We then read the data into R

```
# Load the data and assign it to a variable
file   = "repdata-data-StormData.csv.bz2"
raw    =  read.csv(file, stringsAsFactors = FALSE) # FALSE to optimize read speed
```

We summarize information about the data using the `str` command:

```
str(raw)
```

```
## 'data.frame':    902297 obs. of  37 variables:
##  $ STATE__    : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ BGN_DATE   : chr  "4/18/1950 0:00:00" "4/18/1950 0:00:00" "2/20/1951 0:00:00" "6/8/1951 0:00:00" .
##  $ BGN_TIME   : chr  "0130" "0145" "1600" "0900" ...
##  $ TIME_ZONE : chr  "CST" "CST" "CST" "CST" ...
```

```
##  $ COUNTY    : num  97 3 57 89 43 77 9 123 125 57 ...
##  $ COUNTYNAME: chr  "MOBILE" "BALDWIN" "FAYETTE" "MADISON" ...
##  $ STATE     : chr  "AL" "AL" "AL" "AL" ...
##  $ EVTYPE    : chr  "TORNADO" "TORNADO" "TORNADO" "TORNADO" ...
##  $ BGN_RANGE : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BGN_AZI   : chr  "" "" "" "" ...
##  $ BGN_LOCATI: chr  "" "" "" "" ...
##  $ END_DATE  : chr  "" "" "" "" ...
##  $ END_TIME  : chr  "" "" "" "" ...
##  $ 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   : chr  "" "" "" "" ...
##  $ END_LOCATI: chr  "" "" "" "" ...
##  $ 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: chr  "K" "K" "K" "K" ...
##  $ CROPDMG   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CROPDMGEXP: chr  "" "" "" "" ...
##  $ WFO       : chr  "" "" "" "" ...
##  $ STATEOFFIC: chr  "" "" "" "" ...
##  $ ZONENAMES : chr  "" "" "" "" ...
##  $ 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   : chr  "" "" "" "" ...
##  $ REFNUM    : num  1 2 3 4 5 6 7 8 9 10 ...
```

Looking at the summary of data below, we identify the variables of interest for the analysis. This will be `EVTYPE` (the event type), `FATALITIES`, `INJURIES`, `PROPDMG` (monetary estimate of property damage), and `PROPDMGEXP` (unit used for the damage estimate). These variables either need to be converted or manipulated into more workable formats:

```
  # reformat data type of key variables
  raw$EVTYPE = as.factor(raw$EVTYPE)
  raw$BGN_DATE = as.POSIXlt(strptime(raw$BGN_DATE,format="%m/%d/%Y %H:%M:%S"))
  raw$DMG = ""
  raw$DMG = mapvalues(raw$PROPDMGEXP, c("B","M","m","K","H","h","O"),
                                      c(1e9,1e6,1e6,1e3,1e2,1e2,1))
  raw$DMG = as.numeric(raw$DMG) * raw$PROPDMG
```

```
## Warning: NAs introduced by coercion
```

```
  raw$DMG = as.numeric(raw$DMG)
```

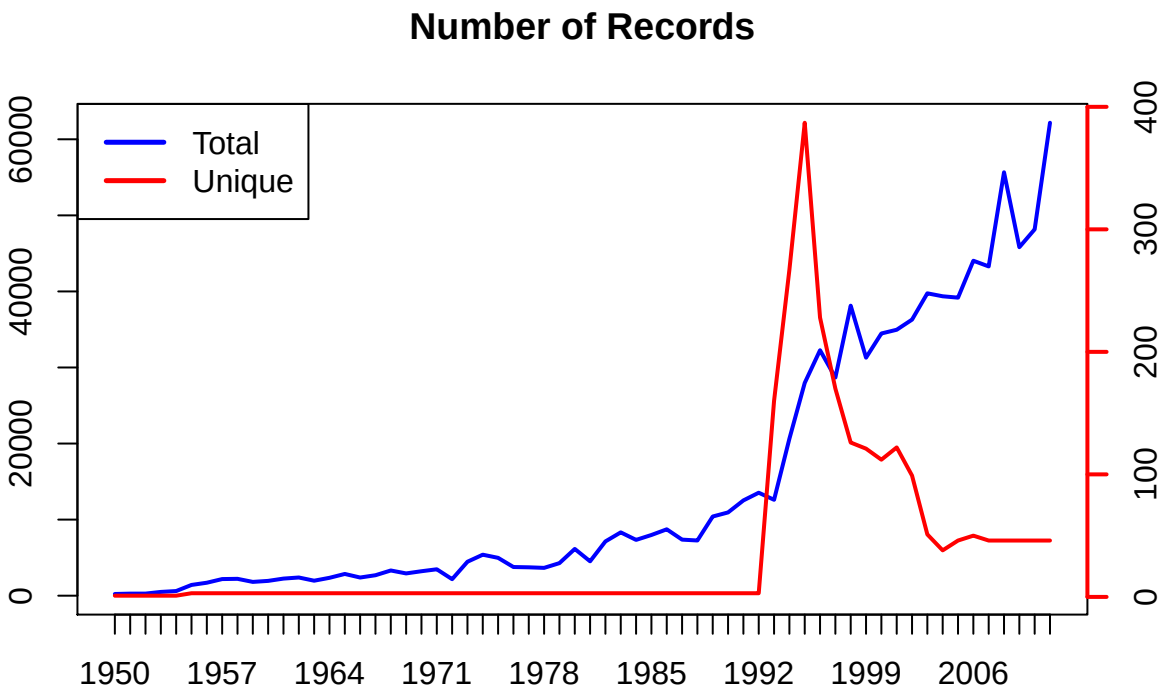
A new variable `DMG` is created to capture the monetary estimate of damages from weather events in a universal unit of measure. Although there are certain uncaught response types that cause NAs to be coerced, these

cases are ambigous to interpret(even after reading the data help files). Luckily, they are few enough that they can likely be ignored without making a big difference on the exploratory analysis.

By plotting the number of *unique* types of weather events per year below, we can see that the initial period of data (~1950 to 1995) has few categorizations. I find it more likely that this absence of data is a result of lack of collection systems/standards, rather than an absence of particular types of events. It is likely that including this initial period would bias the analysis away from type of events that only started being tracked recently.

```
t = table(format(raw$BGN_DATE,"%Y"))
u  = as.numeric(tapply(raw$EVTYPE,raw$BGN_DATE[[6]], function(x) length(unique(x))))

plot(t, type = "l", main = "Number of Records", ylab = "", col = "blue")
par(new=T); plot(u, type='l', col = "red", lwd = 2, axes=F, xlab=NA, ylab=NA)
axis(4, col = "red", lwd = 2)
legend("topleft", c("Total","Unique"), lwd=c(2.5,2.5), col=c("blue","red"))
```



The number of unique weather events jumps sharply in 1995 (387 records). Accordingly, I will work only with the subset of data from this date forward, as it is more likely more representative, and will not bias the data towards type of weather events that were tracked earlier on in history.

```
df = raw[raw$BGN_DATE >= 1995,]
```

This captures 96% of the raw data.

Results

On the basis of fatalities, it appears that XXX is the most harmful to population health.

On the basis of damage

Discuss results

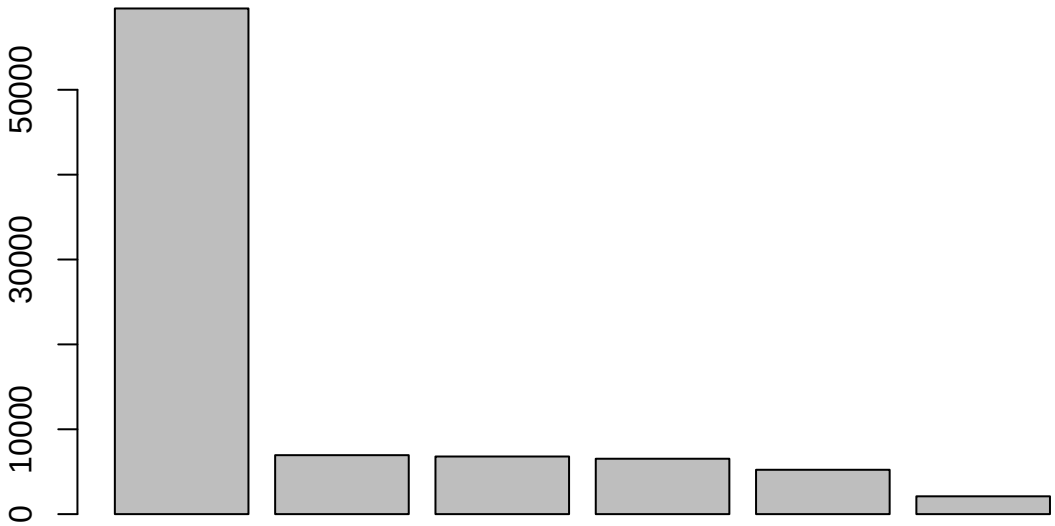
```r
df1 = aggregate(FATALITIES ~ EVTYPE, data = df, sum)
df1 = df1[order(df1$FATALITIES, decreasing = T),]
head(df1)
```

```
##             EVTYPE FATALITIES
## 833        TORNADO       3272
## 130 EXCESSIVE HEAT       1903
## 153    FLASH FLOOD        974
## 275           HEAT        937
## 463      LIGHTNING        812
## 855      TSTM WIND        504
```

```r
df2 = aggregate(INJURIES ~ EVTYPE, data = df, sum)
df2 = df2[order(df2$INJURIES, decreasing = T),]
head(df2)
```

```
##             EVTYPE INJURIES
## 833        TORNADO    59580
## 855      TSTM WIND     6947
## 170          FLOOD     6789
## 130 EXCESSIVE HEAT     6525
## 463      LIGHTNING     5226
## 275           HEAT     2100
```

```r
barplot(head(df2$INJURIES))
```

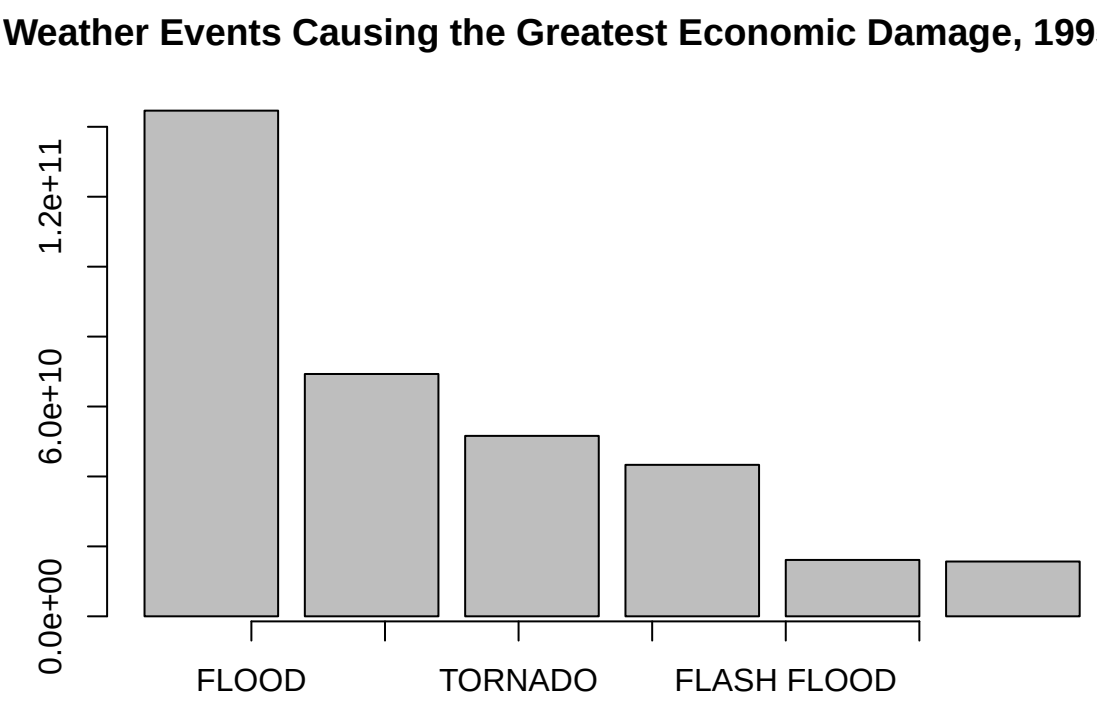


Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health?

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

Property damage estimates

```
df3 = aggregate(DMG ~ EVTYPE, data = df, sum)
df3 = df3[order(df3$DMG, decreasing = T),]
barplot(head(df3$DMG), main = "Weather Events Causing the Greatest Economic Damage, 1995-2008")
axis(1, at = 1:6, labels = head(df3$EVTYPE))
```



Session Info

```
sessionInfo()
```

```
## R version 3.1.2 (2014-10-31)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
##
## locale:
## [1] en_CA.UTF-8/en_CA.UTF-8/en_CA.UTF-8/C/en_CA.UTF-8/en_CA.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] plyr_1.8.1 knitr_1.8
##
## loaded via a namespace (and not attached):
##  [1] codetools_0.2-9  digest_0.6.6     evaluate_0.5.5   formatR_1.0
##  [5] htmltools_0.2.6  Rcpp_0.11.3      rmarkdown_0.3.10 stringr_0.6.2
##  [9] tools_3.1.2      yaml_2.1.13
```