Reproducible Research: Peer Assessment 2

The basic goal of this assignment is to explore the NOAA Storm Database and answer some 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?

Data Processing

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

```
library(knitr)
opts_chunk$set(echo=TRUE)       ## set global parameter for echo
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")) {
    message("Downloading data...")
    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)
```

```
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 ...
```

```
  # 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$PROPDMGEXP = as.factor(raw$PROPDMGEXP)
```

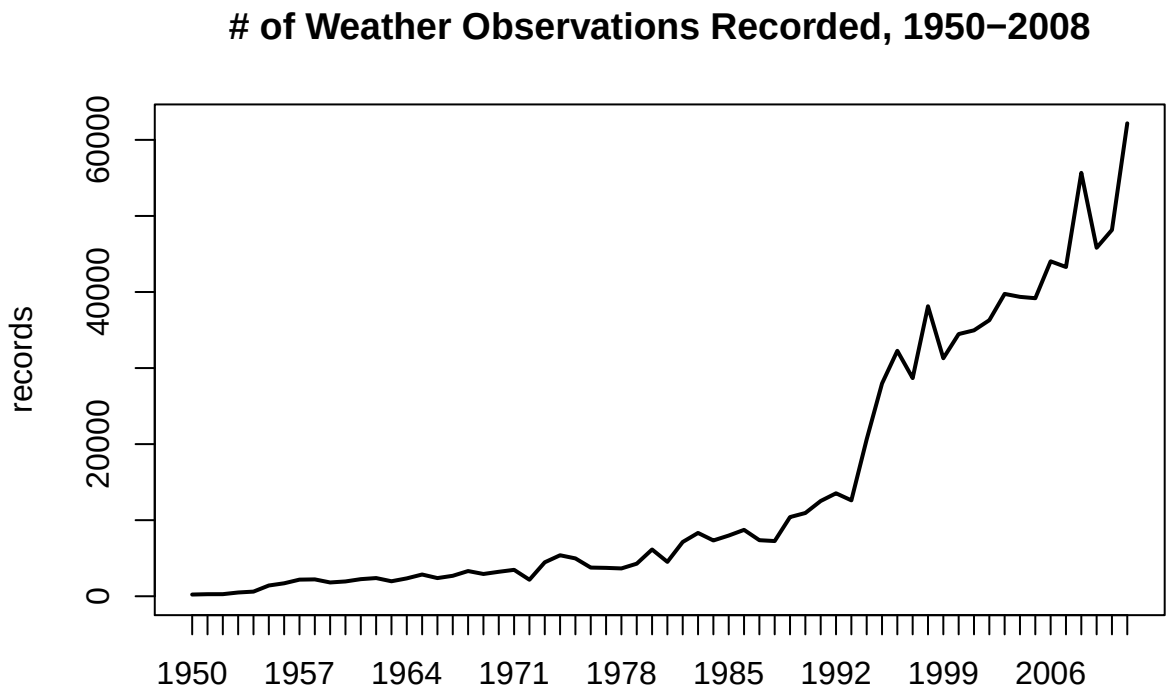
Synopsis

The purpose of the analysis is to determine which types of events are most harmful with respect to population health in the United States.

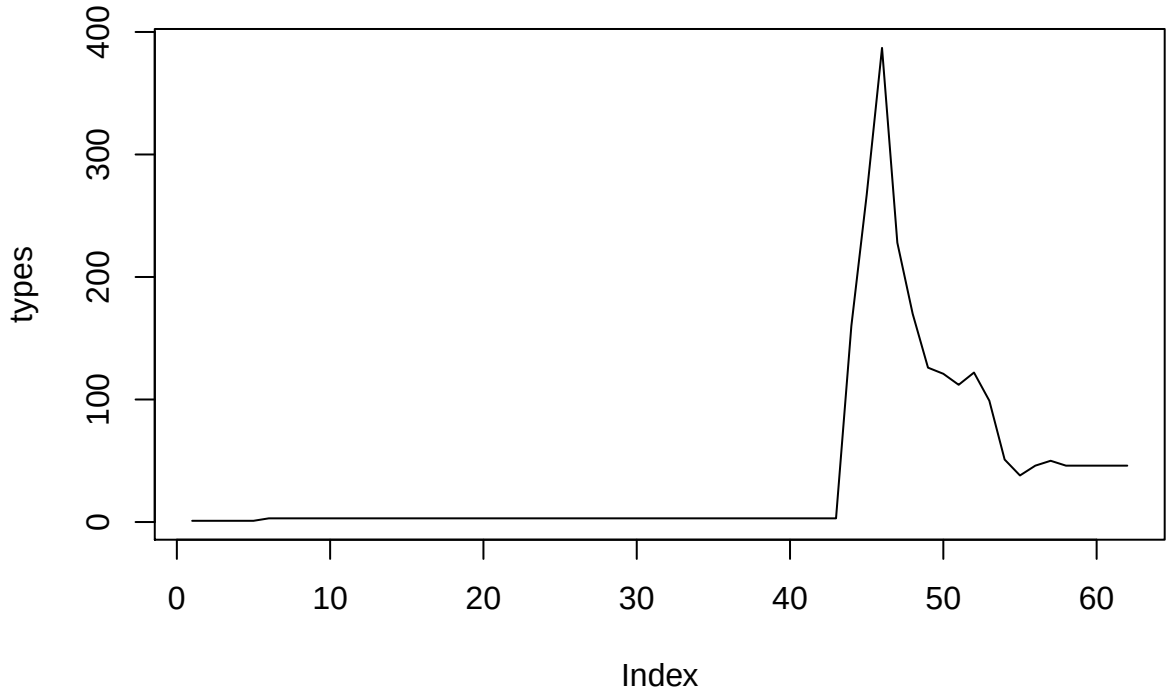
Results

Discuss results

```
records = table(format(raw$BGN_DATE,"%Y"))
plot(records, type = "l", main = "# of Weather Observations Recorded, 1950-2008")
```




```
types = tapply(raw$EVTYPE,raw$BGN_DATE[[6]], function(x) length(unique(x)))
plot(types,type="l")
```




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

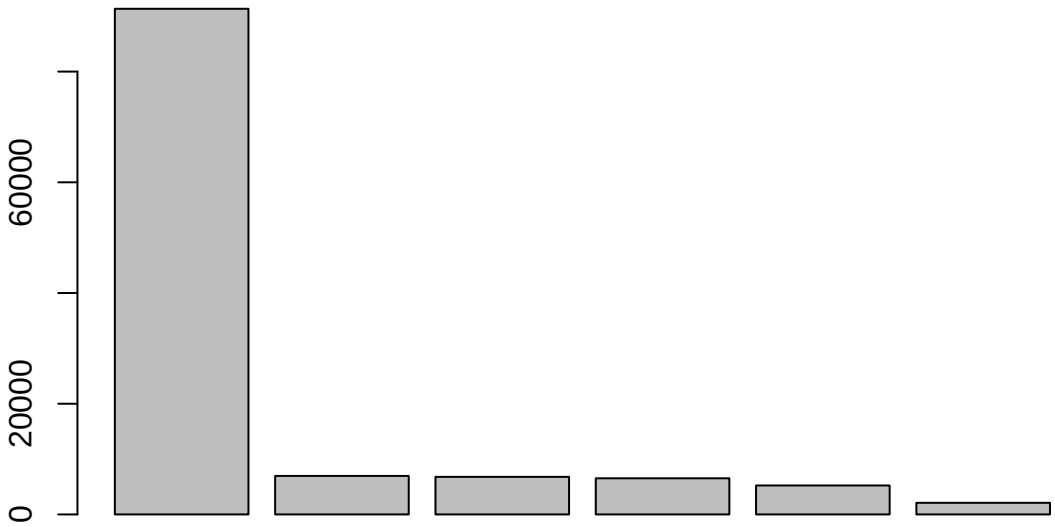
```
##             EVTYPE FATALITIES
```

```
## 834        TORNADO       5633
## 130 EXCESSIVE HEAT       1903
## 153    FLASH FLOOD        978
## 275           HEAT        937
## 464      LIGHTNING        816
## 856      TSTM WIND        504
```

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

```
##             EVTYPE INJURIES
## 834        TORNADO    91346
## 856      TSTM WIND     6957
## 170          FLOOD     6789
## 130 EXCESSIVE HEAT     6525
## 464      LIGHTNING     5230
## 275           HEAT     2100
```

```
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

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] 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  rmarkdown_0.3.10 stringr_0.6.2    tools_3.1.2
## [9] yaml_2.1.13
```