Reproducible Research: Peer Assessment 2

Impact of Severe Weather Events on Public Health and Economy in the United States

xzw0005

Saturday, July 18, 2015

Synopsis

In this report, we aim to analyze the impact of different severe weather events on population health as well as economy across the United States. By exploring the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database, we conclude that across the United States,:

- **Excessive heat**, **tornado** and **flood** are the most harmful severe weather events with respect to population health, and
- **Flood**, **drought** and **hurricane/typhoon** are among the events which have the greatest economic consequences.

Data Processing

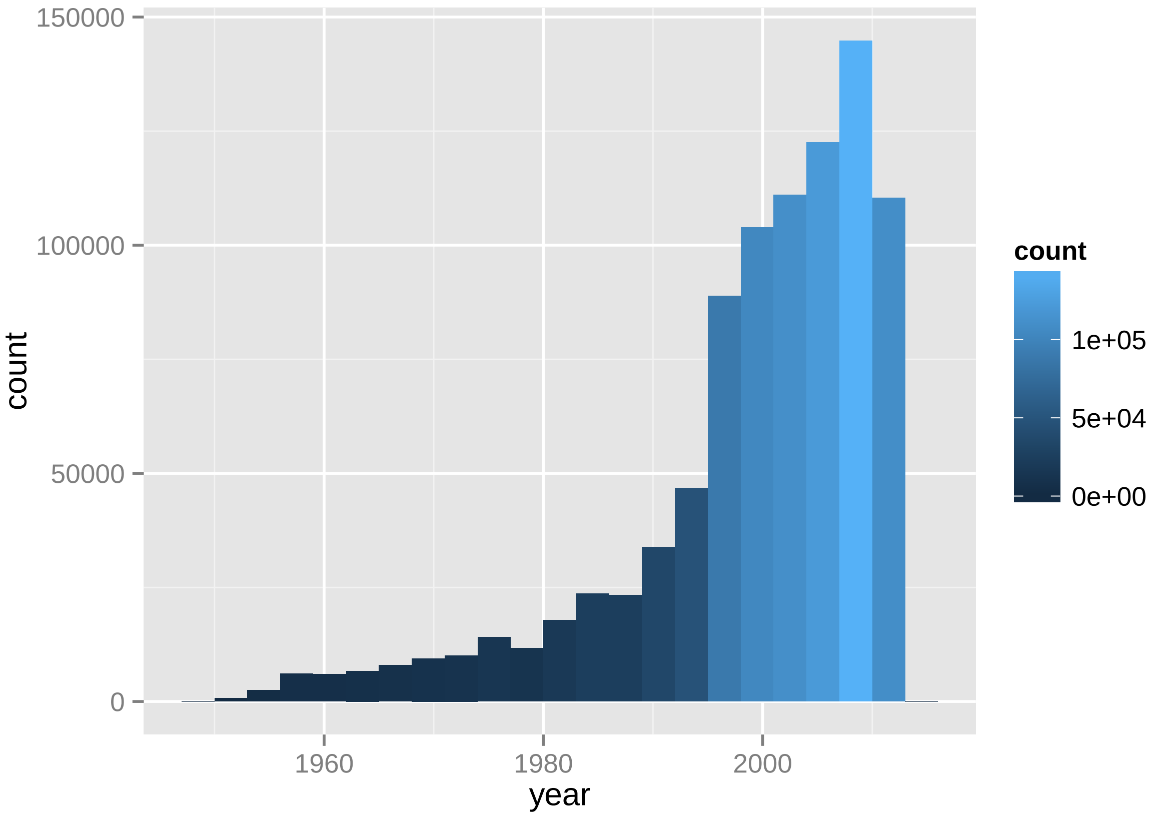
The data comes from the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database. This database tracks characteristics of major storms and weather events in the United States, including when and where they occur, as well as estimates of any fatalities, injuries, and property damage.

```
if (! "StormData" %in% ls()) {
  if (!file.exists("repdata-data-StormData.csv.bz2")) {
    download.file("https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2", destfile
  }
  StormData = read.csv(bzfile("repdata-data-StormData.csv.bz2"))
}
str(StormData)
```

```
## 'data.frame':    902297 obs. of  37 variables:
##  $ STATE__   : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ BGN_DATE  : Factor w/ 16335 levels "1/1/1966 0:00:00",..: 6523 6523 4242 11116 2224 2224 2260 383
##  $ BGN_TIME  : Factor w/ 3608 levels "00:00:00 AM",..: 272 287 2705 1683 2584 3186 242 1683 3186 318
##  $ TIME_ZONE : Factor w/ 22 levels "ADT","AKS","AST",..: 7 7 7 7 7 7 7 7 7 7 ...
##  $ COUNTY    : num  97 3 57 89 43 77 9 123 125 57 ...
##  $ COUNTYNAME: Factor w/ 29601 levels "","5NM E OF MACKINAC BRIDGE TO PRESQUE ISLE LT MI",..: 13513
##  $ STATE     : Factor w/ 72 levels "AK","AL","AM",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ EVTYPE    : Factor w/ 985 levels "   HIGH SURF ADVISORY",..: 834 834 834 834 834 834 834 834 834
##  $ BGN_RANGE : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BGN_AZI   : Factor w/ 35 levels "","  N"," NW",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ BGN_LOCATI: Factor w/ 54429 levels "","- 1 N Albion",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ END_DATE  : Factor w/ 6663 levels "","1/1/1993 0:00:00",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ END_TIME  : Factor w/ 3647 levels ""," 0900CST",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ 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   : Factor w/ 24 levels "","E","ENE","ESE",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ END_LOCATI: Factor w/ 34506 levels "","- .5 NNW",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ 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: 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 ...
##  $ WFO       : Factor w/ 542 levels ""," CI","$AC",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ STATEOFFIC: Factor w/ 250 levels "","ALABAMA, Central",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ ZONENAMES : Factor w/ 25112 levels "","
##  $ 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   : Factor w/ 436781 levels "","-2 at Deer Park\n",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ REFNUM    : num  1 2 3 4 5 6 7 8 9 10 ...
```

There are 902297 observations with 37 variables in total. The events in the database start in the year 1950 and end in November 2011. In the earlier years of the database there are generally fewer events recorded, most likely due to a lack of good records.

```r
StormData$year <- as.numeric(format(as.Date(StormData$BGN_DATE, format = "%m/%d/%Y %H:%M:%S"), "%Y"))
library(ggplot2)
g = ggplot(StormData, aes(x = year))
g + geom_histogram(binwidth = 3, aes(fill = ..count..))
```



According to the histogram over time, we see that the recorded data significantly increased from the year around 1995. Since more recent years should be considered more complete, we would use the subset of the data since the year 1995.

In this project, we mainly focus on the impact of weather events on population health and economy, thus choose 7 relevant variables, they are:

- **EVTYPE:** Type of event
- **FATALITIES:** Number of fatalities
- **INJURIES:** Number of injuries
- **PROPDMG:** Amount of property damage in orders of magnitude and hence economic damage in USD
- **PROPDMGEXP:** Order of magnitude for property damage (e.g. K for thousands)
- **CROPDMG:** Amount of crop damage in orders of magnitude and hence economic damage in USD
- **CROPDMGEXP:** Order of magnitude for crop damage (e.g. M for millions)

```
mydata = StormData[StormData$year >= 1995, c("EVTYPE", "FATALITIES", "INJURIES", "PROPDMG", "PROPDMGEXP
str(mydata)
```

```
## 'data.frame':    681500 obs. of  7 variables:
##  $ EVTYPE    : Factor w/ 985 levels "   HIGH SURF ADVISORY",..: 201 629 657 657 410 244 786 786 244
##  $ FATALITIES: num  0 0 0 0 2 0 0 0 0 0 ...
##  $ INJURIES  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ PROPDMG   : num  0 0 0 0 0.1 0 0 0 0 0 ...
##  $ PROPDMGEXP: Factor w/ 19 levels "","-","?","+",..: 1 1 1 1 14 1 1 1 1 1 ...
##  $ CROPDMG   : num  0 0 0 0 10 0 0 0 0 0 ...
##  $ CROPDMGEXP: Factor w/ 9 levels "","?","0","2",..: 1 1 1 1 9 1 1 1 1 1 ...
```

Therefore, we have obtained the much smaller dataset that we will use. The data contains 681500 observations, although the starting year becomes 1995, it still contains more that 75% of the observations. Note that there are 985 factor levels of different event types in the dataset. However, the National Weather Service Storm Data Documentation describes only 48 types of events (see Section 7, "Event Types", P18-P92). Thus, the event types worth paying much attention. By carefully inspection of the data, we notice that there are several reasons cause so many levels, such as abbreviation, misspelling, different expression, etc. For example, the "FLOOD" might be abbreviated as "FLD", the "FOG" could be misspelled as "VOG", the "STORM" might be misspelled as "STROM", the "AVALANCHE" might be misspelled as "AVALANCE", the winter weather might be expressed as "WINTER" OR "WINTRY" the "THUNDER" might be spelled as "THUNDER", "THUNDERE", etc. At beginning, we are trying to solve this problem by using regular expression. However, the types are quite complicated to deal with. For example, "Marine Thunderstorm Wind" and "Thunderstorm Wind" are totally different types of weather events, where the former is with designator Marine, and the latter is defined as County/Parish (i.e., local). So dealing with the event type names would be our future work.

Another thing needs to be take care of before doing the analysis is that we must take into account the order of magnitude of property damage and crop damage.

```
unique(mydata$PROPDMGEXP)
```

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

```
unique(mydata$CROPDMGEXP)
```

```
## [1]   M m K B ? 0 k 2
## Levels:  ? 0 2 B k K m M
```

Let H = hundreds, K = thousands, M = millions, B = billions. Then we could obtain the damage values for both property and crop.

```
mydata$PROPDMGEXP = as.character(mydata$PROPDMGEXP)
mydata$PROPDMGEXP[mydata$PROPDMGEXP == NA] = 0
mydata$PROPDMGEXP[mydata$PROPDMGEXP == "B"] = 9
mydata$PROPDMGEXP[mydata$PROPDMGEXP == "M"] = 6
mydata$PROPDMGEXP[mydata$PROPDMGEXP == "m"] = 6
mydata$PROPDMGEXP[mydata$PROPDMGEXP == "K"] = 3
mydata$PROPDMGEXP[mydata$PROPDMGEXP == "h"] = 2
mydata$PROPDMGEXP[mydata$PROPDMGEXP == "H"] = 2
mydata$PROPDMGEXP[mydata$PROPDMGEXP == ""] = 0
mydata$PROPDMGEXP[mydata$PROPDMGEXP == "+"] = 0
mydata$PROPDMGEXP[mydata$PROPDMGEXP == "-"] = 0
mydata$PROPDMGEXP[mydata$PROPDMGEXP == "?"] = 0

mydata$PropDmgVal = mydata$PROPDMG * (10 ^ as.numeric(mydata$PROPDMGEXP))
```

```r
mydata$CROPDMGEXP = as.character(mydata$CROPDMGEXP)
mydata$CROPDMGEXP[mydata$CROPDMGEXP == NA] = 0
mydata$CROPDMGEXP[mydata$CROPDMGEXP == ""] = 0
mydata$CROPDMGEXP[mydata$CROPDMGEXP == "?"] = 0
mydata$CROPDMGEXP[mydata$CROPDMGEXP == "B"] = 9
mydata$CROPDMGEXP[mydata$CROPDMGEXP == "M"] = 6
mydata$CROPDMGEXP[mydata$CROPDMGEXP == "m"] = 6
mydata$CROPDMGEXP[mydata$CROPDMGEXP == "K"] = 3
str(mydata)
```

```
## 'data.frame':    681500 obs. of  8 variables:
##  $ EVTYPE    : Factor w/ 985 levels "   HIGH SURF ADVISORY",..: 201 629 657 657 410 244 786 786 244
##  $ FATALITIES: num  0 0 0 0 2 0 0 0 0 0 ...
##  $ INJURIES  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ PROPDMG   : num  0 0 0 0 0.1 0 0 0 0 0 ...
##  $ PROPDMGEXP: chr  "0" "0" "0" "0" ...
##  $ CROPDMG   : num  0 0 0 0 10 0 0 0 0 0 ...
##  $ CROPDMGEXP: chr  "0" "0" "0" "0" ...
##  $ PropDmgVal: num  0e+00 0e+00 0e+00 0e+00 1e+08 0e+00 0e+00 0e+00 0e+00 0e+00 ...
```

```r
mydata$CropDmgVal = mydata$CROPDMG * (10 ^ as.numeric(mydata$CROPDMGEXP))
```

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

Results

Most harmful events with respect to population health

To begin our analysis for the most harmful events with respect to population health, we are about to aggregate the number of fatalities and injuries by the type of severe weather events.

```r
#fatalities = aggregate(mydata$FATALITIES, by = list(mydata$EVTYPE), FUN = "sum")
fatalities = aggregate(FATALITIES ~ EVTYPE, data = mydata, FUN = sum)
fatalities = fatalities[order(-fatalities$FATALITIES), ]
head(fatalities)
```

```
##             EVTYPE FATALITIES
## 112 EXCESSIVE HEAT       1903
## 666        TORNADO       1545
## 134    FLASH FLOOD        934
## 231           HEAT        924
## 358      LIGHTNING        729
## 144          FLOOD        423
```

```r
injuries = aggregate(INJURIES ~ EVTYPE, data = mydata, FUN = sum)
injuries = injuries[order(-injuries$INJURIES), ]
head(injuries)
```

```
##             EVTYPE INJURIES
## 666        TORNADO    21765
## 144          FLOOD     6769
## 112 EXCESSIVE HEAT     6525
## 358      LIGHTNING     4631
## 683      TSTM WIND     3630
## 231           HEAT     2030
```

We can see from the above the top six severe weather events causing the largest number of fatalities and injuries, respectively. Now, let's visualize it via barplots.

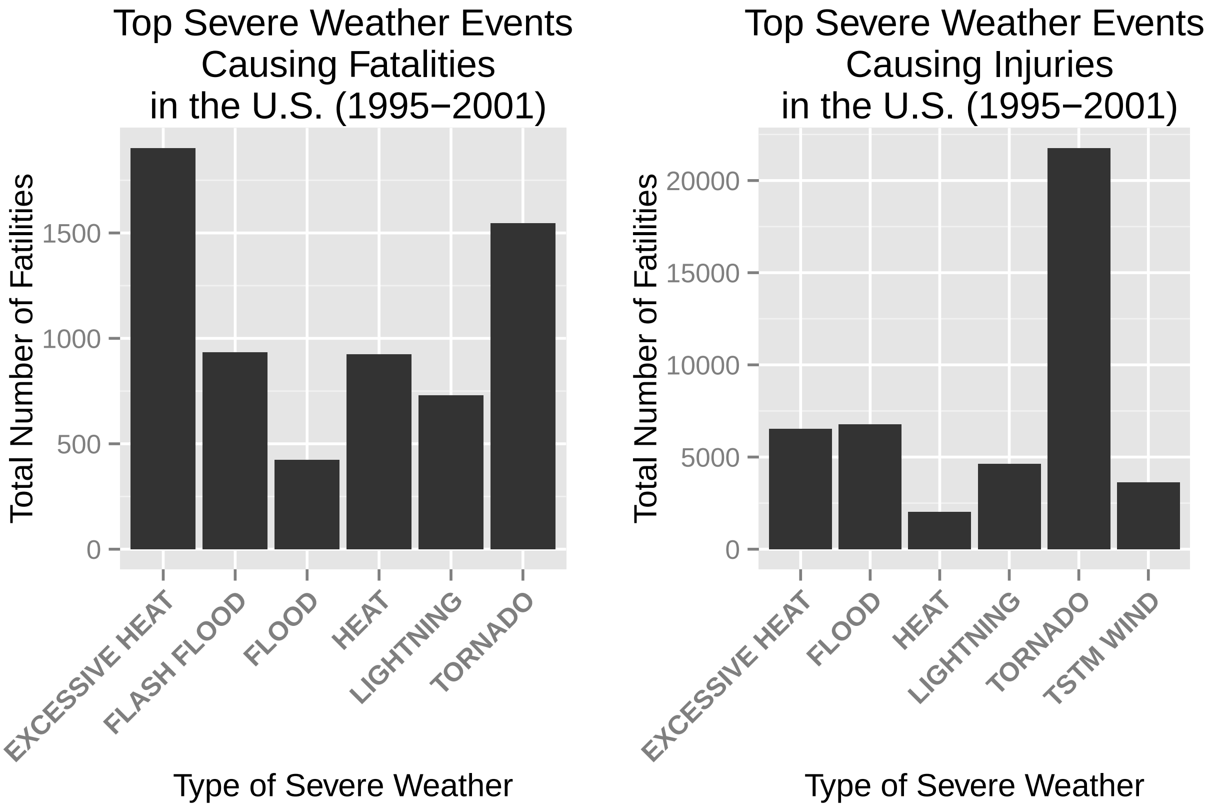
```
library(ggplot2)
library(gridExtra)
```

```
## Loading required package: grid
```

```
fatalPlot = ggplot(data = head(fatalities), aes(x = EVTYPE, y = FATALITIES))
fatalPlot = fatalPlot + geom_bar(stat="identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold")) +
  ggtitle("Top Severe Weather Events\n Causing Fatalities\n in the U.S. (1995-2001)") +
  xlab("Type of Severe Weather") + ylab("Total Number of Fatilities")

injurePlot = ggplot(data = head(injuries), aes(x = EVTYPE, y = INJURIES)) + geom_bar(stat="identity")
injurePlot = injurePlot + theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold")) +
  ggtitle("Top Severe Weather Events\n Causing Injuries\n in the U.S. (1995-2001)") +
  xlab("Type of Severe Weather") + ylab("Total Number of Fatilities")

grid.arrange(fatalPlot, injurePlot, ncol=2)
```



According to the above pair of histograms, we could see that **excessive heat** and **tornado** cause most *fatalities*, while **tornato** and **flood** cause most *injuries* in the United States from 1995 to 2011. Therefore, we conclude that **excessive heat**, **tornado** and **flood** are the most harmful severe weather events with respect to population health.

Which types of events have the greatest economic consequences

To begin our analysis for the event types have the greatest economic consequences, we are about to aggregate the value (in U.S. $) of damage and loss by the type of severe weather events.

```r
propertyDamage = aggregate(PropDmgVal ~ EVTYPE, data = mydata, FUN = sum)
propertyDamage = propertyDamage[order(-propertyDamage$PropDmgVal), ]
head(propertyDamage)
```

```
##                EVTYPE   PropDmgVal
## 144             FLOOD 144022037057
## 313 HURRICANE/TYPHOON  69305840000
## 519       STORM SURGE  43193536000
## 666           TORNADO  24935939545
## 134       FLASH FLOOD  16047794571
## 206              HAIL  15048722103
```

```r
cropDamage = aggregate(CropDmgVal ~ EVTYPE, data = mydata, FUN = sum)
cropDamage = cropDamage[order(-cropDamage$CropDmgVal), ]
head(cropDamage)
```

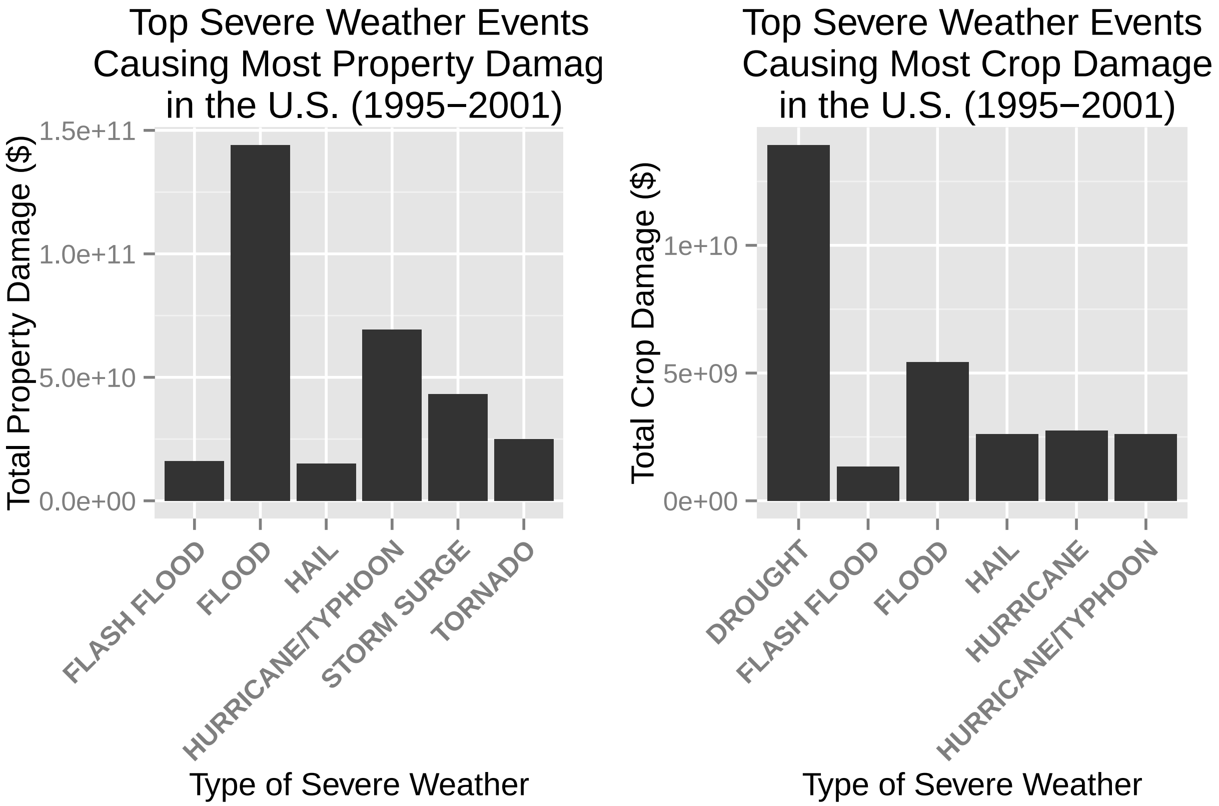
```
##                EVTYPE  CropDmgVal
## 84            DROUGHT 13922066000
## 144             FLOOD  5422810400
## 306         HURRICANE  2741410000
## 206              HAIL  2613777070
## 313 HURRICANE/TYPHOON  2607872800
## 134       FLASH FLOOD  1343915000
```

We can see from the above the top six severe weather events causing the most economic damage on property and crop, respectively. Now, let's visualize it via barplots.

```r
propPlot = ggplot(data = head(propertyDamage), aes(x = EVTYPE, y = PropDmgVal))
propPlot = propPlot + geom_bar(stat="identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold")) +
  ggtitle("Top Severe Weather Events\n Causing Most Property Damage \n in the U.S. (1995-2001)") +
  xlab("Type of Severe Weather") + ylab("Total Property Damage ($)")

cropPlot = ggplot(data = head(cropDamage), aes(x = EVTYPE, y = CropDmgVal)) + geom_bar(stat="identity")
cropPlot = cropPlot + theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold")) +
  ggtitle("Top Severe Weather Events\n Causing Most Crop Damage\n in the U.S. (1995-2001)") +
  xlab("Type of Severe Weather") + ylab("Total Crop Damage ($)")

grid.arrange(propPlot, cropPlot, ncol=2)
```



According to the above pair of histograms, we could see that that **flood** and **hurricane/typhoon** cause most *property damage*, while **drought** and **flood** cause most *crop damage* in the United States from 1995 to 2011. Therefore, we conclude that **flood**, **drought** and **hurricane/typhoon** are among the events which have the greatest economic consequences.