Severe Weather Analysis

June 20, 2017

Synopsys

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. This analysis is built to answer two questions:

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

The results show that Tornados caused the most fatalities and injuries. However, floods and heat-related events caused a lot of fatalities as well.

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

Different events caused different economic damages. Heat-related events were the most damaging to crops, while floods caused the most property damage.

Data Processing

Downloading the data from the source and loading it into R

The data was available course web site. Format: comma-separated-value file compressed via the bzip2 algorithm Read.csv is able to read this format without any additional parameters specified This step can take a few minutes depending on computer

```
download.file("https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2",
              "repdata2Fdata2FStormData.csv.bz2")

stormData <- read.csv("repdata2Fdata2FStormData.csv.bz2")
```

Selecting the analysis variables

Taking a look at the structure of the file

```
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 3980 3980 ...
##  $ BGN_TIME  : Factor w/ 3608 levels "00:00:00 AM",..: 272 287 2705 1683 2584 3186 242 1683 3186 3186 ...
##  $ 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 1873 4598 10592 4372
 10094 1973 23873 24418 4598 ...
##  $ 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",..: 834 834 834 834 834 834 834 834 834 834 ...
##  $ 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 "","                                                                                                                               
                                               "| __truncated__,..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ 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 ...
```

To speed up processing, we create a subset with variables of interest.

1. To determine event types most harmful with respect to population health we will use “FATALITIES” and “INJURIES”
2. To determine event types with the greatest economic consequences we will use “PROPDMG”, “PROPDMGEXP”, “CROPDMG”, “CROPDMGEXP”

```
analysisVars <- c("EVTYPE", "FATALITIES", "INJURIES", "PROPDMG", "PROPDMGEXP", "CROPDMG", "CROPDMGEXP")
stormDataSS <- stormData[analysisVars]
summary(stormDataSS)
```

```
##                EVTYPE         FATALITIES          INJURIES       
##  HAIL             :288661   Min.   :  0.0000   Min.   :   0.0000  
##  TSTM WIND        :219940   1st Qu.:  0.0000   1st Qu.:   0.0000  
##  THUNDERSTORM WIND: 82563   Median :  0.0000   Median :   0.0000  
##  TORNADO          : 60652   Mean   :  0.0168   Mean   :   0.1557  
##  FLASH FLOOD      : 54277   3rd Qu.:  0.0000   3rd Qu.:   0.0000  
##  FLOOD            : 25326   Max.   :583.0000   Max.   :1700.0000  
##  (Other)          :170878                                        
##     PROPDMG          PROPDMGEXP        CROPDMG          CROPDMGEXP    
##  Min.   :   0.00          :465934   Min.   :  0.000          :618413  
##  1st Qu.:   0.00   K      :424665   1st Qu.:  0.000   K      :281832  
##  Median :   0.00   M      : 11330   Median :  0.000   M      :  1994  
##  Mean   :  12.06   0      :   216   Mean   :  1.527   k      :    21  
##  3rd Qu.:   0.50   B      :    40   3rd Qu.:  0.000   0      :    19  
##  Max.   :5000.00   5      :    28   Max.   :990.000   B      :     9  
##                    (Other):    84                     (Other):     9  
```

Are there any rows with missing data?

```
sum(is.na(stormDataSS))
```

```
## [1] 0
```

According to “National Weather Service Instruction” document provided with the course, page 12, Damage estimates “should be rounded to three significant digits, followed by an alphabetical character signifying the magnitude of the number, i.e., 1.55B for $1,550,000,000. Alphabetical characters used to signify magnitude include”K" for thousands, “M” for millions, and “B” for billions."

- $ Estimates are contained in PROPDMG and CROPDMG variables (both numeric)
- Magnitude of the estimates is contained in PROPDMGEXP and CROPDMGEXP (both factor).

However, from the summary above, it appears that these two factor variables contain levels not defined in the description.

Let’s check the details

```
table(stormDataSS$PROPDMGEXP)
```

```
## 
##             -      ?      +      0      1      2      3      4      5 
## 465934      1      8      5    216     25     13      4      4     28 
##      6      7      8      B      h      H      K      m      M 
##      4      5      1     40      1      6 424665      7  11330 
```

```
table(stormDataSS$CROPDMGEXP)
```

```
## 
##             ?      0      2      B      k      K      m      M 
## 618413      7     19      1      9     21 281832      1   1994 
```

We will assume that any levels outside specified (K, M, and B) indicate the magnitude of $1. Let’s recode these variables accordingly:

```
stormDataSS$PROPDMGEXP_R <- 1
stormDataSS$PROPDMGEXP_R[stormDataSS$PROPDMGEXP == 'B'] <- 1000000000
stormDataSS$PROPDMGEXP_R[stormDataSS$PROPDMGEXP == 'b'] <- 1000000000
stormDataSS$PROPDMGEXP_R[stormDataSS$PROPDMGEXP == 'K'] <- 1000
stormDataSS$PROPDMGEXP_R[stormDataSS$PROPDMGEXP == 'k'] <- 1000
stormDataSS$PROPDMGEXP_R[stormDataSS$PROPDMGEXP == 'M'] <- 1000000
stormDataSS$PROPDMGEXP_R[stormDataSS$PROPDMGEXP == 'm'] <- 1000000

stormDataSS$CROPDMGEXP_R <- 1
stormDataSS$CROPDMGEXP_R[stormDataSS$CROPDMGEXP == 'B'] <- 1000000000
stormDataSS$CROPDMGEXP_R[stormDataSS$CROPDMGEXP == 'b'] <- 1000000000
stormDataSS$CROPDMGEXP_R[stormDataSS$CROPDMGEXP == 'K'] <- 1000
stormDataSS$CROPDMGEXP_R[stormDataSS$CROPDMGEXP == 'k'] <- 1000
stormDataSS$CROPDMGEXP_R[stormDataSS$CROPDMGEXP == 'M'] <- 1000000
stormDataSS$CROPDMGEXP_R[stormDataSS$CROPDMGEXP == 'm'] <- 1000000
```

Then we will transform the actual estimates based on their magnitude. Given that some numbers are very large, we will convert all estimated into millions (MM)

```
stormDataSS$PROPDMG_MM <- round(stormDataSS$PROPDMG * stormDataSS$PROPDMGEXP_R / 1000000)
stormDataSS$CROPDMG_MM <- round(stormDataSS$CROPDMG * stormDataSS$CROPDMGEXP_R / 1000000)
```

Let's check how the numbers sum up

1.Property damage

```
tapply(stormDataSS$PROPDMG_MM, stormDataSS$PROPDMGEXP_R, sum)
```

```
##      1   1000  1e+06  1e+09
##      0   1801 139657 275850
```

2. Crop damage

```
tapply(stormDataSS$CROPDMG_MM, stormDataSS$CROPDMGEXP_R, sum)
```

```
##     1  1000 1e+06 1e+09
##     0   245 34143 13610
```

From the numbers above it appears that Property Damage was much higher ($275B) compared to Crop Damage ($13.6B)

Aggregating the data

The analysis needs to be done across all the US and across the years. Hence, we need to aggregate the data by EVTYPE only

```
stormDataSSAggr <- aggregate(cbind(FATALITIES, INJURIES, PROPDMG_MM, CROPDMG_MM) ~ EVTYPE, data = stormDataSS,sum)
summary(stormDataSSAggr)
```

```
##                    EVTYPE      FATALITIES         INJURIES
##     HIGH SURF ADVISORY:  1   Min.   :   0.00   Min.   :    0.0
##   COASTAL FLOOD       :  1   1st Qu.:   0.00   1st Qu.:    0.0
##   FLASH FLOOD         :  1   Median :   0.00   Median :    0.0
##   LIGHTNING           :  1   Mean   :  15.38   Mean   :  142.7
##   TSTM WIND           :  1   3rd Qu.:   0.00   3rd Qu.:    0.0
##   TSTM WIND (G45)     :  1   Max.   :5633.00   Max.   :91346.0
##  (Other)              :979
##    PROPDMG_MM         CROPDMG_MM
##  Min.   :     0.0   Min.   :    0.00
##  1st Qu.:     0.0   1st Qu.:    0.00
##  Median :     0.0   Median :    0.00
##  Mean   :   423.7   Mean   :   48.73
##  3rd Qu.:     0.0   3rd Qu.:    0.00
##  Max.   :144123.0   Max.   :13957.00
##
```

Based on the above information, there are a lot of events that don't have any fatalities, injuries or damage. Let's create variables that indicate no harm to people's health or no economical damage

1. No harm to people's health:

```
stormDataSSAggr$noHarm <- (stormDataSSAggr$FATALITIES + stormDataSSAggr$INJURIES) == 0
table(stormDataSSAggr$noHarm)
```

```
##
## FALSE  TRUE
##   220   765
```

Hence, there were 765 event types that caused on harm to people's health

2. No damage to property or crops:

```
stormDataSSAggr$noDamage <- (stormDataSSAggr$PROPDMG_MM + stormDataSSAggr$CROPDMG_MM) == 0
table(stormDataSSAggr$noDamage)
```

```
##
## FALSE  TRUE
##   149   836
```

Hence, there were 836 event types that caused on harm to people's health

What about events that caused neither harm nor damage:

```
table(stormDataSSAggr$noDamage, stormDataSSAggr$noHarm)
```

```
##
##         FALSE TRUE
##   FALSE    87   62
##   TRUE    133  703
```

703 events caused neither harm nor damage. We can remove them from the analysis

```
stormDataSSAggr <- subset(stormDataSSAggr, (noDamage + noHarm) < 2)
table(stormDataSSAggr$noDamage, stormDataSSAggr$noHarm)
```

```
##
##         FALSE TRUE
##   FALSE    87   62
##   TRUE    133    0
```

This is the dataset we will use for the analysis

Results

Determining types of events that were most harmful with respect to population health

Let's first look at the distribution of Fatalities

```
quantile(stormDataSSAggr$FATALITIES, probs = c(seq(0, 1, by = 0.1)))
```

```
##     0%    10%    20%    30%    40%    50%    60%    70%    80%    90%
##    0.0    0.0    0.0    0.0    0.0    1.0    2.0    3.7    9.8   60.7
##   100%
## 5633.0
```

It appears that we can focus on events that contain top 10% of fatalities

```
quantile(stormDataSSAggr$INJURIES, probs = c(seq(0, 1, by = 0.1)))
```

```
##      0%     10%     20%     30%     40%     50%     60%     70%     80%
##     0.0     0.0     0.0     0.0     0.0     1.0     2.0     8.0    30.6
##     90%    100%
##   231.9 91346.0
```

Same applies to injuries

We will create variables that split events into top 10% vs. rest for both fatalities and injuries

```
stormDataSSAggr$Top10F <- cut(stormDataSSAggr$FATALITIES,
                              breaks = c(quantile(stormDataSSAggr$FATALITIES, probs = c(0, 0.9, 1))),
                              labels = c("0-90","90-100"))
stormDataSSAggr$Top10I <- cut(stormDataSSAggr$INJURIES,
                              breaks = c(quantile(stormDataSSAggr$INJURIES, probs = c(0, 0.9, 1))),
                              labels = c("0-90","90-100"))
```

How do these new variables look?

```
table(stormDataSSAggr$Top10F)
```

```
##
##   0-90 90-100
##    139     29
```

```
tapply(stormDataSSAggr$FATALITIES, stormDataSSAggr$Top10F, sum)
```

```
##   0-90 90-100
##    928  14217
```

```
table(stormDataSSAggr$Top10I)
```

```
##
##   0-90 90-100
##    129     29
```

```
tapply(stormDataSSAggr$INJURIES, stormDataSSAggr$Top10I, sum)
```

```
##   0-90 90-100
##   3132 137396
```

There are 29 event types in both top 10% by fatalities and injuries. How big is the overlap?

```
table(stormDataSSAggr$Top10F, stormDataSSAggr$Top10I)
```

```
##         
##          0-90 90-100
##   0-90     70      7
##   90-100    7     22
```

22 event types were the most damaging in terms of public health by both injuries and fatalities. Before looking at them, let’s sort our dataset by fatalities in reverse order

```
stormDataSSAggrSF <- stormDataSSAggr[order(-stormDataSSAggr$FATALITIES),]
```

Now let’s create a dataset with only those 22 types of events and look at it

```
stormDataSSAggrSF <- subset(stormDataSSAggrSF, Top10F == '90-100' & Top10I == '90-100', select=c(EVTYPE, FATALITIES, INJURIE
S))
stormDataSSAggrSF$EVTYPE <- factor(stormDataSSAggrSF$EVTYPE)
stormDataSSAggrSF
```

```
##                EVTYPE FATALITIES INJURIES
## 834           TORNADO       5633    91346
## 130    EXCESSIVE HEAT       1903     6525
## 153       FLASH FLOOD        978     1777
## 275              HEAT        937     2100
## 464         LIGHTNING        816     5230
## 856         TSTM WIND        504     6957
## 170             FLOOD        470     6789
## 585       RIP CURRENT        368      232
## 359         HIGH WIND        248     1137
## 972      WINTER STORM        206     1321
## 586      RIP CURRENTS        204      297
## 278         HEAT WAVE        172      309
## 760  THUNDERSTORM WIND        133     1488
## 310        HEAVY SNOW        127     1021
## 676       STRONG WIND        103      280
## 30           BLIZZARD        101      805
## 290        HEAVY RAIN         98      251
## 427         ICE STORM         89     1975
## 957          WILDFIRE         75      911
## 411  HURRICANE/TYPHOON         64     1275
## 786 THUNDERSTORM WINDS         64      908
## 188               FOG         62      734
```

Based on the table above that the most harmful type of event in the US is Tornado. It caused the most fatalities and injuries. Some of the event types, however, should be grouped since they are very similar (i.e., Heat and Heat wave).

```
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'DENSE FOG'] <- 'Rain/Fog'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'FOG'] <- 'Rain/Fog'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'HAIL'] <- 'Rain/Fog'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'HEAVY RAIN'] <- 'Rain/Fog'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'LIGHTNING'] <- 'Rain/Fog'

stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'WILD/FOREST FIRE'] <- 'Wildfire'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'WILDFIRE'] <- 'Wildfire'

stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'AVALANCHE'] <- 'Avalanche'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'AVALANCE'] <- 'Avalanche'

stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'BLIZZARD'] <- 'Cold'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'HEAVY SNOW'] <- 'Cold'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'WINTER STORM'] <- 'Cold'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'EXTREME COLD'] <- 'Cold'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'EXTREME COLD/WIND CHILL'] <- 'Cold'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'ICE STORM'] <- 'Cold'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'WINTER WEATHER'] <- 'Cold'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'AGRICULTURAL FREEZE'] <- 'Cold'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'BLACK ICE'] <- 'Cold'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'blowing snow'] <- 'Cold'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'BLOWING SNOW'] <- 'Cold'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'Cold'] <- 'Cold'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'COLD AND WET CONDITIONS'] <- 'Cold'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'Cold Temperature'] <- 'Cold'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'COLD WAVE'] <- 'Cold'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'COLD WEATHER'] <- 'Cold'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'COLD/WIND CHILL'] <- 'Cold'

stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'DROUGHT'] <- 'Heat'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'EXCESSIVE HEAT'] <- 'Heat'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'HEAT'] <- 'Heat'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'HEAT WAVE'] <- 'Heat'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'EXTREME HEAT'] <- 'Heat'

stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'FLASH FLOOD'] <- 'Flood'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'FLOOD'] <- 'Flood'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'RIVER FLOOD'] <- 'Flood'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'HIGH SURF'] <- 'Flood'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'RIP CURRENT'] <- 'Flood'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'RIP CURRENTS'] <- 'Flood'

stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'ASTRONOMICAL HIGH TIDE'] <- 'Coastal events'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'COASTAL  FLOODING/EROSION'] <- ''
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'COASTAL EROSION'] <- 'Coastal events'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'Coastal Flood'] <- 'Coastal events'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'Coastal Flooding'] <- 'Coastal events'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'COASTAL FLOODING'] <- 'Coastal events'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'Coastal Flooding'] <- 'Coastal events'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'Coastal Flooding'] <- 'Coastal events'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'Coastal Storm'] <- 'Coastal events'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'COASTAL STORM'] <- 'Coastal events'

stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'HIGH WIND'] <- 'Storm/Wind'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'TSTM WIND'] <- 'Storm/Wind'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'STRONG WIND'] <- 'Storm/Wind'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'THUNDERSTORM WIND'] <- 'Storm/Wind'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'THUNDERSTORM WINDS'] <- 'Storm/Wind'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'STORM SURGE'] <- 'Storm/Wind'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'STORM SURGE/TIDE'] <- 'Storm/Wind'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'TROPICAL STORM'] <- 'Storm/Wind'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == ' TSTM WIND'] <- 'Storm/Wind'

stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'HURRICANE'] <- 'Hurricane'
stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'HURRICANE/TYPHOON'] <- 'Hurricane'

stormDataSSAggrSF$EVTYPE_GROUP[stormDataSSAggrSF$EVTYPE == 'TORNADO'] <- 'Tornado'

stormDataSSAggrSF$EVTYPE_GROUP <- factor(stormDataSSAggrSF$EVTYPE_GROUP)
```

The final step is to chart the most harmful events to compare fatalities and injuries Let's create a dataset that can be used for that.

```
stormDataSSAggrSF2 <- aggregate(cbind(FATALITIES, INJURIES) ~ EVTYPE_GROUP, data = stormDataSSAggrSF,sum)
```

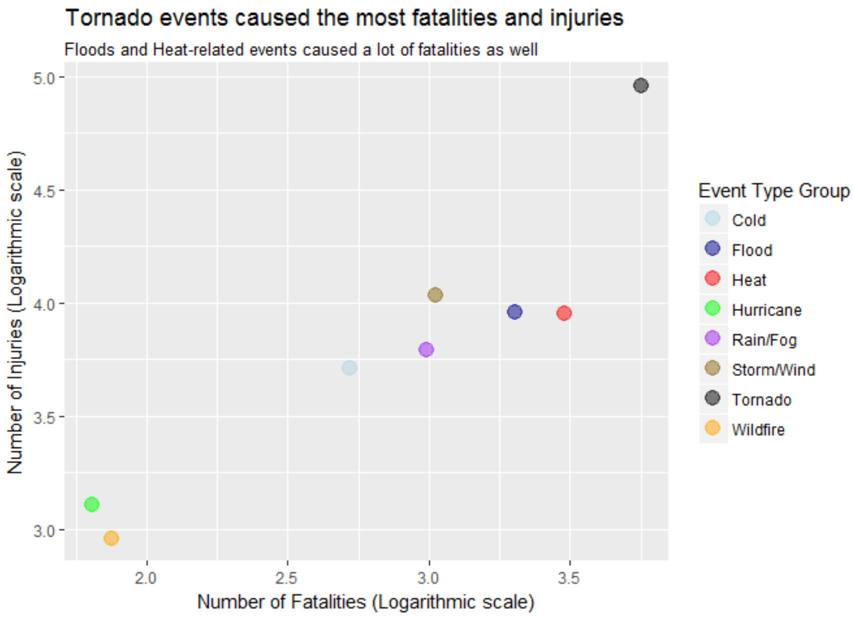
There is a large variance in fatalities and injuries across event types. Therefore, we'll use logarithmic scale for the chart

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
cols <- c("Cold" = "lightblue",
          "Heat" = "red",
          "Hurricane" = "green",
          "Rain/Fog" = "purple",
          "Storm/Wind" = "goldenrod4",
          "Tornado" = "black",
          "Flood" = "darkblue",
          "Wildfire" = "orange")

g1 <- ggplot(stormDataSSAggrSF2, aes(log10(FATALITIES), log10(INJURIES), color = EVTYPE_GROUP))
g1 +
    geom_point(size = 4, alpha = 1/2) +
    scale_color_manual(values=cols) +
    labs(color = "Event Type Group") +
    ggtitle("Tornado events caused the most fatalities and injuries",
            subtitle = "Floods and Heat-related events caused a lot of fatalities as well") +
    xlab("Number of Fatalities (Logarithmic scale)") +
    ylab("Number of Injuries (Logarithmic scale)")
```



Determining types of events have the greatest economic consequences

```
quantile(stormDataSSAggr$PROPDMG_MM, probs = c(seq(0, 1, by = 0.1)))
```

```
##       0%      10%      20%      30%      40%      50%      60%      70% 
##      0.0      0.0      0.0      0.0      0.0      0.0      1.0      5.0 
##      80%      90%     100% 
##     19.2    538.3 144123.0
```

It appears that we can focus on events that contain top 10% of property damages

```
quantile(stormDataSSAggr$CROPDMG_MM, probs = c(seq(0, 1, by = 0.1)))
```

```
##      0%     10%     20%     30%     40%     50%     60%     70%     80% 
##     0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0     6.0 
##     90%    100% 
##    91.2 13957.0
```

Same applies to crop damages

We will create variables that split events into top 10% vs. rest for both fatalities and injuries

```
stormDataSSAggr$Top10PD <- cut(stormDataSSAggr$PROPDMG_MM,
                              breaks = c(quantile(stormDataSSAggr$PROPDMG_MM, probs = c(0, 0.9, 1))),
                              labels = c("0-90","90-100"))
stormDataSSAggr$Top10CD <- cut(stormDataSSAggr$CROPDMG_MM,
                              breaks = c(quantile(stormDataSSAggr$CROPDMG_MM, probs = c(0, 0.9, 1))),
                              labels = c("0-90","90-100"))
```

How do these new variables look?

```
table(stormDataSSAggr$Top10PD)
```

```
##
##   0-90 90-100
##    101     29
```

```
tapply(stormDataSSAggr$PROPDMG_MM, stormDataSSAggr$Top10PD, sum)
```

```
##   0-90 90-100
##   4006 413302
```

```
table(stormDataSSAggr$Top10CD)
```

```
##
##   0-90 90-100
##     43     29
```

```
tapply(stormDataSSAggr$CROPDMG_MM, stormDataSSAggr$Top10CD, sum)
```

```
##   0-90 90-100
##    836  47162
```

There are 29 event types in both top 10% by property and crop damages. How big is the overlap?

```
table(stormDataSSAggr$Top10PD, stormDataSSAggr$Top10CD)
```

```
##
##          0-90 90-100
##   0-90     21      7
##   90-100    6     19
```

19 event types were the most damaging in terms of economic consequences. Before looking at them, let's combine them into one variable and sort our dataset by the new variable

```
stormDataSSAggr$TTLDMG_MM <- stormDataSSAggr$PROPDMG_MM + stormDataSSAggr$CROPDMG_MM
stormDataSSAggrSDMG <- stormDataSSAggr[order(-stormDataSSAggr$TTLDMG_MM),]
```

Now let's create a dataset with only those 29 types of events and look at it

```
stormDataSSAggrSDMG <- subset(stormDataSSAggrSDMG, Top10PD == '90-100' & Top10CD == '90-100', select=c(EVTYPE, PROPDMG_MM, C
ROPDMG_MM, TTLDMG_MM))
stormDataSSAggrSDMG$EVTYPE <- factor(stormDataSSAggrSDMG$EVTYPE)
stormDataSSAggrSDMG
```

```
##                EVTYPE PROPDMG_MM CROPDMG_MM TTLDMG_MM
## 170             FLOOD     144123       5535    149658
## 411  HURRICANE/TYPHOON      69308       2607     71915
## 834           TORNADO      53063        334     53397
## 244              HAIL      15177       2534     17711
## 153        FLASH FLOOD      15028       1283     16311
## 95            DROUGHT       1045      13957     15002
## 402          HURRICANE      11864       2740     14604
## 590        RIVER FLOOD       5112       5026     10138
## 427          ICE STORM       3897       5021      8918
## 848     TROPICAL STORM       7676        673      8349
## 359          HIGH WIND       5047        629      5676
## 957           WILDFIRE       4723        289      5012
## 856          TSTM WIND       3310        461      3771
## 955   WILD/FOREST FIRE       2983        105      3088
## 760  THUNDERSTORM WIND       2698        360      3058
## 786 THUNDERSTORM WINDS       1300        172      1472
## 290         HEAVY RAIN        663        726      1389
## 310         HEAVY SNOW        841        134       975
## 30            BLIZZARD        639        112       751
```

Based on the table above that the most damaging in terms of economic impact type of event in the US is Flood, It caused the most combined (property and crop damages). Some of the event types, however, should be grouped since they are very similar (i.e., Flood and Flash Flood).

```
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'DENSE FOG'] <- 'Rain/Fog'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'FOG'] <- 'Rain/Fog'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'HAIL'] <- 'Rain/Fog'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'HEAVY RAIN'] <- 'Rain/Fog'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'LIGHTNING'] <- 'Rain/Fog'

stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'WILD/FOREST FIRE'] <- 'Wildfire'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'WILDFIRE'] <- 'Wildfire'

stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'AVALANCHE'] <- 'Avalanche'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'AVALANCE'] <- 'Avalanche'

stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'BLIZZARD'] <- 'Cold'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'HEAVY SNOW'] <- 'Cold'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'WINTER STORM'] <- 'Cold'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'EXTREME COLD'] <- 'Cold'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'EXTREME COLD/WIND CHILL'] <- 'Cold'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'ICE STORM'] <- 'Cold'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'WINTER WEATHER'] <- 'Cold'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'AGRICULTURAL FREEZE'] <- 'Cold'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'BLACK ICE'] <- 'Cold'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'blowing snow'] <- 'Cold'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'BLOWING SNOW'] <- 'Cold'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'Cold'] <- 'Cold'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'COLD AND WET CONDITIONS'] <- 'Cold'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'Cold Temperature'] <- 'Cold'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'COLD WAVE'] <- 'Cold'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'COLD WEATHER'] <- 'Cold'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'COLD/WIND CHILL'] <- 'Cold'

stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'DROUGHT'] <- 'Heat'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'EXCESSIVE HEAT'] <- 'Heat'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'HEAT'] <- 'Heat'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'HEAT WAVE'] <- 'Heat'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'EXTREME HEAT'] <- 'Heat'

stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'FLASH FLOOD'] <- 'Flood'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'FLOOD'] <- 'Flood'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'RIVER FLOOD'] <- 'Flood'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'HIGH SURF'] <- 'Flood'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'RIP CURRENT'] <- 'Flood'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'RIP CURRENTS'] <- 'Flood'

stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'ASTRONOMICAL HIGH TIDE'] <- 'Coastal events'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'COASTAL  FLOODING/EROSION'] <- ''
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'COASTAL EROSION'] <- 'Coastal events'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'Coastal Flood'] <- 'Coastal events'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'Coastal Flooding'] <- 'Coastal events'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'COASTAL FLOODING'] <- 'Coastal events'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'Coastal Flooding'] <- 'Coastal events'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'Coastal Flooding'] <- 'Coastal events'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'Coastal Storm'] <- 'Coastal events'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'COASTAL STORM'] <- 'Coastal events'

stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'HIGH WIND'] <- 'Storm/Wind'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'TSTM WIND'] <- 'Storm/Wind'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'STRONG WIND'] <- 'Storm/Wind'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'THUNDERSTORM WIND'] <- 'Storm/Wind'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'THUNDERSTORM WINDS'] <- 'Storm/Wind'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'STORM SURGE'] <- 'Storm/Wind'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'STORM SURGE/TIDE'] <- 'Storm/Wind'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'TROPICAL STORM'] <- 'Storm/Wind'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == ' TSTM WIND'] <- 'Storm/Wind'

stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'HURRICANE'] <- 'Hurricane'
stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'HURRICANE/TYPHOON'] <- 'Hurricane'

stormDataSSAggrSDMG$EVTYPE_GROUP[stormDataSSAggrSDMG$EVTYPE == 'TORNADO'] <- 'Tornado'

stormDataSSAggrSDMG$EVTYPE_GROUP <- factor(stormDataSSAggrSDMG$EVTYPE_GROUP)
```

The final step is to chart the most harmful events to compare property and crop damages Let’s create a dataset that can be used for that.

```
stormDataSSAggrSDMG2 <- aggregate(cbind(PROPDMG_MM, CROPDMG_MM) ~ EVTYPE_GROUP, data = stormDataSSAggrSDMG,sum)
```

There is a large variance in fatalities and injuries across event types. Therefore, we’ll use logarithmic scale for the chart

```
g2 <- ggplot(stormDataSSAggrSDMG2, aes(PROPDMG_MM, CROPDMG_MM, color = EVTYPE_GROUP))
g2 +
    geom_point(size = 4, alpha = 1/2) +
    scale_color_manual(values=cols) +
    labs(color = "Event Type Group") +
    ggtitle("Heat caused the most property damages while floods damaged crops the most") +
    xlab("Property Damages ($MM)") +
    ylab("Crop Damages ($MM)")
```