

# Public Health and Economic Impact of Weather events in the US

## Synopsis

Here

## Data Processing

1. **Note 1:** Dependencies: no
2. **Note 2:** The source documentation for this analysis is given in [NWSI](#)

For our analysis, we are going to use the NOAA Storm Database. So first we need to download it to a temporal file, expand it and put it in a data frame called `weather_dataset`:

```
filename = "http://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
tempfile <- tempfile()
download.file(filename, tempfile)
weather_dataset = read.csv(bzfile(tempfile), sep=";", header=T)
unlink(tempfile)
```

With the following function let's check if there are any NA's in the dataset:

```
nacols <- function(df) {
  colnames(df)[unlist(lapply(df, function(x) any(is.na(x))))]
}
na_cols = nacols(weather_dataset)
na_cols
```

```
## [1] "COUNTYENDN" "F" "LATITUDE" "LATITUDE_E"
```

As we can see there are 4 columns that contain NA values. So let's keep in mind this just in case we have to use them.

We will also check the number of rows with NA's, to have an idea of the completeness of our dataset:

```
ok = complete.cases(weather_dataset)
na_rows = sum(!ok)
na_rows
```

```
## [1] 902297
```

As we can see, there are a lot (902297) of missing values in this dataset.

Lets get to know a bit out dataset. These are the fields:

```
str(weather_dataset)
```

```
## '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 ...
## $ CROPDGMG : num  0 0 0 0 0 0 0 0 0 0 ...
## $ CROPDGMGEXP: Factor w/ 9 levels "", "?", "0", "2", ...: 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 ...
```

This are the field names:

```
colnames(weather_dataset)
```

```
## [1] "STATE__"      "BGN_DATE"      "BGN_TIME"      "TIME_ZONE"     "COUNTY"
## [6] "COUNTYNAME"  "STATE"         "EVTTYPE"       "BGN_RANGE"     "BGN_AZI"
## [11] "BGN_LOCATI"   "END_DATE"      "END_TIME"      "COUNTY_END"   "COUNTYENDN"
## [16] "END_RANGE"    "END_AZI"       "END_LOCATI"    "LENGTH"        "WIDTH"
## [21] "F"            "MAG"           "FATALITIES"    "INJURIES"      "PROPDMG"
## [26] "PROPDMGEXP"   "CROPDMG"       "CROPDMGEXP"    "WFO"            "STATEOFFIC"
## [31] "ZONENAMES"    "LATITUDE"      "LONGITUDE"     "LATITUDE_E"    "LONGITUDE_"
## [36] "REMARKS"      "REFNUM"
```

and this is a simple summary:

```
summary(weather_dataset)
```

```
##      STATE__      BGN_DATE      BGN_TIME
```

```

## Min.      : 1.0      5/25/2011 0:00:00: 1202      12:00:00 AM: 10163
## 1st Qu.:19.0      4/27/2011 0:00:00: 1193      06:00:00 PM: 7350
## Median :30.0      6/9/2011 0:00:00 : 1030      04:00:00 PM: 7261
## Mean   :31.2      5/30/2004 0:00:00: 1016      05:00:00 PM: 6891
## 3rd Qu.:45.0      4/4/2011 0:00:00 : 1009      12:00:00 PM: 6703
## Max.    :95.0      4/2/2006 0:00:00 : 981       03:00:00 PM: 6700
##
## (Other)      :895866 (Other)      :857229
##
## TIME_ZONE      COUNTY      COUNTYNAME      STATE
## CST      :547493 Min.      : 0      JEFFERSON : 7840 TX      : 83728
## EST      :245558 1st Qu.: 31      WASHINGTON: 7603 KS      : 53440
## MST      : 68390 Median : 75      JACKSON  : 6660 OK      : 46802
## PST      : 28302 Mean   :101      FRANKLIN : 6256 MO      : 35648
## AST      : 6360 3rd Qu.:131      LINCOLN  : 5937 IA      : 31069
## HST      : 2563 Max.    :873      MADISON  : 5632 NE      : 30271
## (Other): 3631      (Other) :862369 (Other):621339
##
## EVTYPE      BGN_RANGE      BGN_AZI
## HAIL      :288661 Min.      : 0      :547332
## TSTM WIND      :219940 1st Qu.: 0      N      : 86752
## THUNDERSTORM WIND: 82563 Median : 0      W      : 38446
## TORNADO      : 60652 Mean   : 1      S      : 37558
## FLASH FLOOD      : 54277 3rd Qu.: 1      E      : 33178
## FLOOD      : 25326 Max.    :3749 NW      : 24041
## (Other)      :170878 (Other):134990
##
## BGN_LOCATI      END_DATE      END_TIME
##      :287743      :243411      :238978
## COUNTYWIDE      : 19680 4/27/2011 0:00:00: 1214 06:00:00 PM: 9802
## Countywide      : 993 5/25/2011 0:00:00: 1196 05:00:00 PM: 8314
## SPRINGFIELD      : 843 6/9/2011 0:00:00 : 1021 04:00:00 PM: 8104
## SOUTH PORTION: 810 4/4/2011 0:00:00 : 1007 12:00:00 PM: 7483
## NORTH PORTION: 784 5/30/2004 0:00:00: 998 11:59:00 PM: 7184
## (Other)      :591444 (Other)      :653450 (Other)      :622432
##
## COUNTY_END COUNTYENDN      END_RANGE      END_AZI
## Min.      :0      Mode:logical Min.      : 0      :724837
## 1st Qu.:0      NA's:902297 1st Qu.: 0      N      : 28082
## Median :0      Median : 0      S      : 22510
## Mean   :0      Mean   : 1      W      : 20119
## 3rd Qu.:0      3rd Qu.: 0      E      : 20047
## Max.    :0      Max.    :925      NE      : 14606
##
## (Other): 72096
##
## END_LOCATI      LENGTH      WIDTH      F

```

```

##          :499225  Min.   : 0.0   Min.   : 0   Min.   :0
## COUNTYWIDE : 19731  1st Qu.: 0.0   1st Qu.: 0   1st Qu.:0
## SOUTH PORTION : 833  Median : 0.0   Median : 0   Median :1
## NORTH PORTION : 780  Mean    : 0.2   Mean    : 8   Mean    :1
## CENTRAL PORTION: 617  3rd Qu.: 0.0   3rd Qu.: 0   3rd Qu.:1
## SPRINGFIELD   : 575  Max.    :2315.0  Max.    :4400  Max.    :5
## (Other)       :380536                      NA's    :843563
##      MAG      FATALITIES    INJURIES      PROPDMG
## Min.   : 0   Min.   : 0   Min.   : 0.0   Min.   : 0
## 1st Qu.: 0   1st Qu.: 0   1st Qu.: 0.0   1st Qu.: 0
## Median : 50  Median : 0   Median : 0.0   Median : 0
## Mean    : 47  Mean    : 0   Mean    : 0.2   Mean    : 12
## 3rd Qu.: 75  3rd Qu.: 0   3rd Qu.: 0.0   3rd Qu.: 0
## Max.    :22000  Max.    :583   Max.    :1700.0  Max.    :5000
##
##      PROPDMGEXP      CROPDGMG      CROPDMGEXP      WFO
##          :465934  Min.   : 0.0          :618413          :142069
## K          :424665  1st Qu.: 0.0   K          :281832   OUN          : 17393
## M          : 11330  Median : 0.0   M          : 1994   JAN          : 13889
## 0          : 216   Mean    : 1.5   k          : 21    LWX          : 13174
## B          : 40    3rd Qu.: 0.0   0          : 19    PHI          : 12551
## 5          : 28    Max.    :990.0  B          : 9     TSA          : 12483
## (Other): 84          (Other): 9   (Other):690738
##
##                      STATEOFFIC
##                      :248769
## TEXAS, North          : 12193
## ARKANSAS, Central and North Central: 11738
## IOWA, Central         : 11345
## KANSAS, Southwest     : 11212
## GEORGIA, North and Central : 11120
## (Other)              :595920
##
ZONENAMES
##
:594029
##
:205988
## GREATER RENO / CARSON CITY / M - GREATER RENO / CARSON CITY / M
: 639
## GREATER LAKE TAHOE AREA - GREATER LAKE TAHOE AREA

```

```

: 592
## JEFFERSON - JEFFERSON
: 303
## MADISON - MADISON
: 302
## (Other)
:100444
## LATITUDE LONGITUDE LATITUDE_E LONGITUDE_
## Min. : 0 Min. : -14451 Min. : 0 Min. : -14455
## 1st Qu.:2802 1st Qu.: 7247 1st Qu.: 0 1st Qu.: 0
## Median :3540 Median : 8707 Median : 0 Median : 0
## Mean :2875 Mean : 6940 Mean :1452 Mean : 3509
## 3rd Qu.:4019 3rd Qu.: 9605 3rd Qu.:3549 3rd Qu.: 8735
## Max. :9706 Max. : 17124 Max. :9706 Max. :106220
## NA's :47 NA's :40
## REMARKS REFNUM
## :287433 Min. : 1
## : 24013 1st Qu.:225575
## Trees down.\n : 1110 Median :451149
## Several trees were blown down.\n : 568 Mean :451149
## Trees were downed.\n : 446 3rd Qu.:676723
## Large trees and power lines were blown down.\n: 432 Max. :902297
## (Other) :588295

```

We have to provide a unique standard unit for the values of crop damage and property damage. I choose K (thousand's of \$)

```

# provide a unique standard unit for prop damage
standarizePropDmgUnit = function(propDmg, propDmgExp) {
  if(propDmgExp=="b") {#billion
    propDmg * 1000 * 1000

  } else if(propDmgExp=="M") {#million
    propDmg * 1000
  } else if(propDmgExp=="m") {#Thousandth
    propDmg / (1000^6)
  } else if(propDmgExp=="H") {#hundred
    propDmg / 1000
  } else {# fr K and all other values, return as is
    #Note: its very obscure the symbol h and the numbers,
    # -, + and ?

```

```

    # I just keep it as is
    propDmg
  }
}

# provides a unique standard unit for
# crop damage values
standardizeCropDmgUnit = function(cropDmg, cropDmgExp) {
  if(cropDmgExp=="b") {#billion
    cropDmg * 1000 * 1000
  } else if(cropDmgExp=="M") {#million
    cropDmg * 1000
  } else if(cropDmgExp=="m") {#Thousandth
    cropDmg / (1000^6)
  } else if(cropDmgExp=="H") {#hundred
    cropDmg / 1000
  } else {# fr K and all other values, return as is
    #Note: its very obscure the symbol h and the numbers,
    # -, + and ?
    # I just keep it as is
    cropDmg
  }
}

#remove entries with 0 fatalities (harmless)
weather_dataset = weather_dataset[weather_dataset$FATALITIES != 0,]

#remove entries with no injuries (harmless)
weather_dataset = weather_dataset[weather_dataset$INJURIES != 0,]

#remove entries with no prop damage expenditures
weather_dataset = weather_dataset[weather_dataset$PROPDMG != 0,]

#remove entries with no crop damage expenditures
weather_dataset = weather_dataset[weather_dataset$CROPDMG != 0,]

for(i in 1:nrow(weather_dataset)) {
  propDmg = weather_dataset[i, "PROPDMG"]
  propDmgExp = weather_dataset[i, "PROPDMGEXP"]
  cropDmg = weather_dataset[i, "CROPDMG"]
}

```

```

cropDmgExp = weather_dataset[i, "CROPDMGEXP"]

weather_dataset[i, "PROPDMG"] = standarizePropDmgUnit(
  propDmg, propDmgExp)
weather_dataset[i, "CROPDMG"] = standarizeCropDmgUnit(
  cropDmg, cropDmgExp)
# do stuff with row
}

# IMPORTANT: please not I didn't "clean" the fields in the sense
# that I didn't merge fields together like others did.
# I tihink that for doing that, one should have more info
# on why those fields that look the same should be merged
# with confidence. else one may be twisting results

```

## Results

We want to answer the following 2 fundamental questions:

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

### Most harmful events for population health

The field for the event types is EVTYPE. Lets take a look at some event types:

```
str(weather_dataset$EVTYPE)
```

```
## Factor w/ 985 levels "    HIGH SURF ADVISORY",...: 834 834 972 973 976 851 170 30 786 30 ...
```

In section 7 of the NWSI reference document ([NWSI](#)) we can inspect the different types of events in detail. For example From that document we can see that Excessive Heat (Z) event type is used for reporting fatalities (directly-related) or major impacts to human health occurring during excessive heat; Here is the quote: **Excessive Heat (Z) Fatalities (directly-related) or major impacts to human health occurring during excessive heat warning conditions are reported using this event category.**



coming back to our question, We need a measure of the impact of each of these event types to public health. From the doc and the dataset, that impact should be given by two fields: FATALITIES (death cases) and INJURIES.

We should treat these 2 in a separate way as they are not the same. But we can “join” them together to see the overall “health impact” as this is what we want to answer

Total deaths per event type (first 10 greatest by # of deaths):

```
deathsPerEvtyp = aggregate(weather_dataset[ 'FATALITIES' ],
  by=list(event = weather_dataset$EVTYPE), FUN=sum)
# order the results
deathsPerEvtyp = deathsPerEvtyp[with(deathsPerEvtyp, order(-FATALITIES)), ]

top10Fatalities = head(deathsPerEvtyp, n=10)
top10Fatalities
```

```
##           event FATALITIES
## 15      TORNADO         190
##  4        FLOOD          58
##  2 EXCESSIVE HEAT          46
## 19        TSUNAMI          32
## 20      WILDFIRE          31
##  3    FLASH FLOOD          23
##  5          HEAT          22
## 11 HURRICANE/TYPHOON        22
##  8      HIGH WIND          15
##  1      BLIZZARD          14
```

Those are the top-10 most deathful events. As we can see, tornados and excessive heat are the most fatal events by far, with 5633 and 1903 number of deaths respectively.

Total injuries per event type (first 10 greatest by # of injuries):

```
injPerEvtyp = aggregate(weather_dataset[ 'INJURIES' ],
  by=list(event = weather_dataset$EVTYPE), FUN=sum)
# order the results
injPerEvtyp = injPerEvtyp[with(injPerEvtyp, order(-INJURIES)), ]

top10Injuries = head(injPerEvtyp, n=10)
```

## top10Injuries

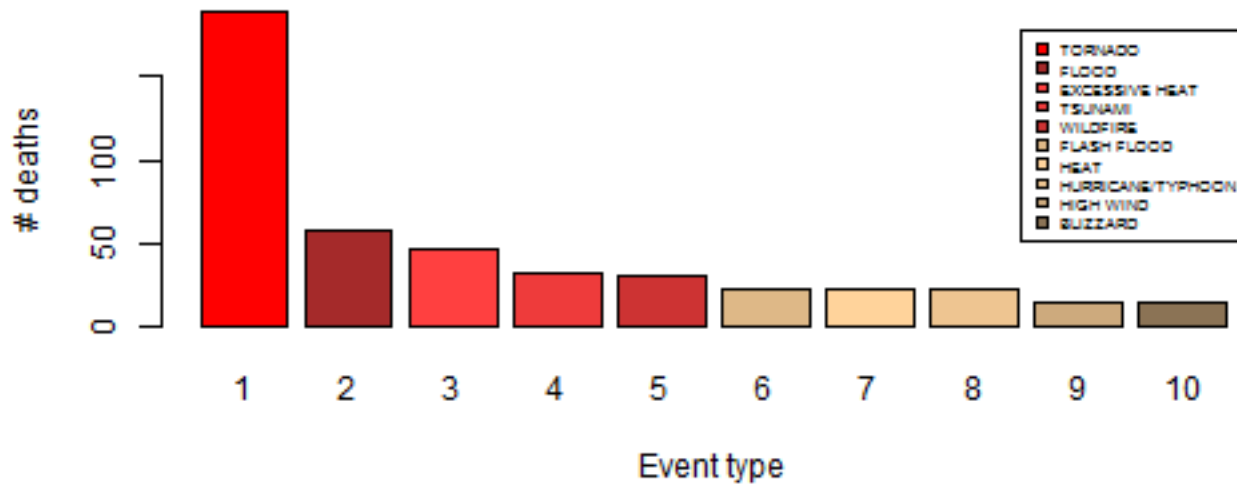
| ##    | event             | INJURIES |
|-------|-------------------|----------|
| ## 4  | FLOOD             | 2495     |
| ## 15 | TORNADO           | 1630     |
| ## 12 | ICE STORM         | 1568     |
| ## 11 | HURRICANE/TYPHOON | 884      |
| ## 1  | BLIZZARD          | 402      |
| ## 5  | HEAT              | 320      |
| ## 16 | TROPICAL STORM    | 267      |
| ## 3  | FLASH FLOOD       | 220      |
| ## 19 | TSUNAMI           | 129      |
| ## 20 | WILDFIRE          | 124      |

Those are the top-10 most harmful (only injuries) events, with tornados been the most harmful (91346 injurie cases) events by far.

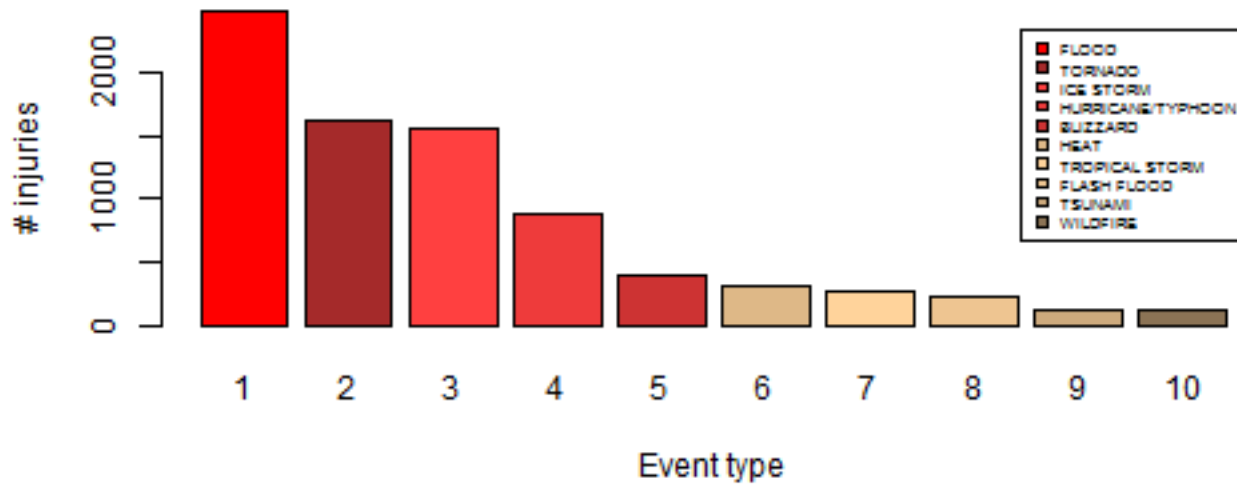
The following is a graph (bar plot) of the top-10 most harmful events in each case (death and injuries):

```
par(mfrow = c(2, 1))
# deaths plot
barplot(top10Fatalities$FATALITIES,main="Top-10 Deaths per event type", xlab="Event type", ylab="#
deaths",col=c("red", "brown", "brown1", "brown2", "brown3", "burlywood", "burlywood1", "burlywood2",
"burlywood3", "burlywood4"), names.arg=1:10, legend=top10Fatalities[, "event"], args.legend=c(cex=0.5))
#args.legend=c(cex=0.4))
# and this is the injuries plot
barplot(top10Injuries$INJURIES,main="Top-10 Injuries count per event type",
        xlab="Event type", ylab="# injuries",col=c("red", "brown", "brown1", "brown2",
"brown3", "burlywood", "burlywood1", "burlywood2", "burlywood3", "burlywood4"), names.arg=1:10,
        legend=top10Injuries[, "event"], args.legend=c(cex=0.5))
```

Top-10 Deaths per event type



Top-10 Injuries count per event type



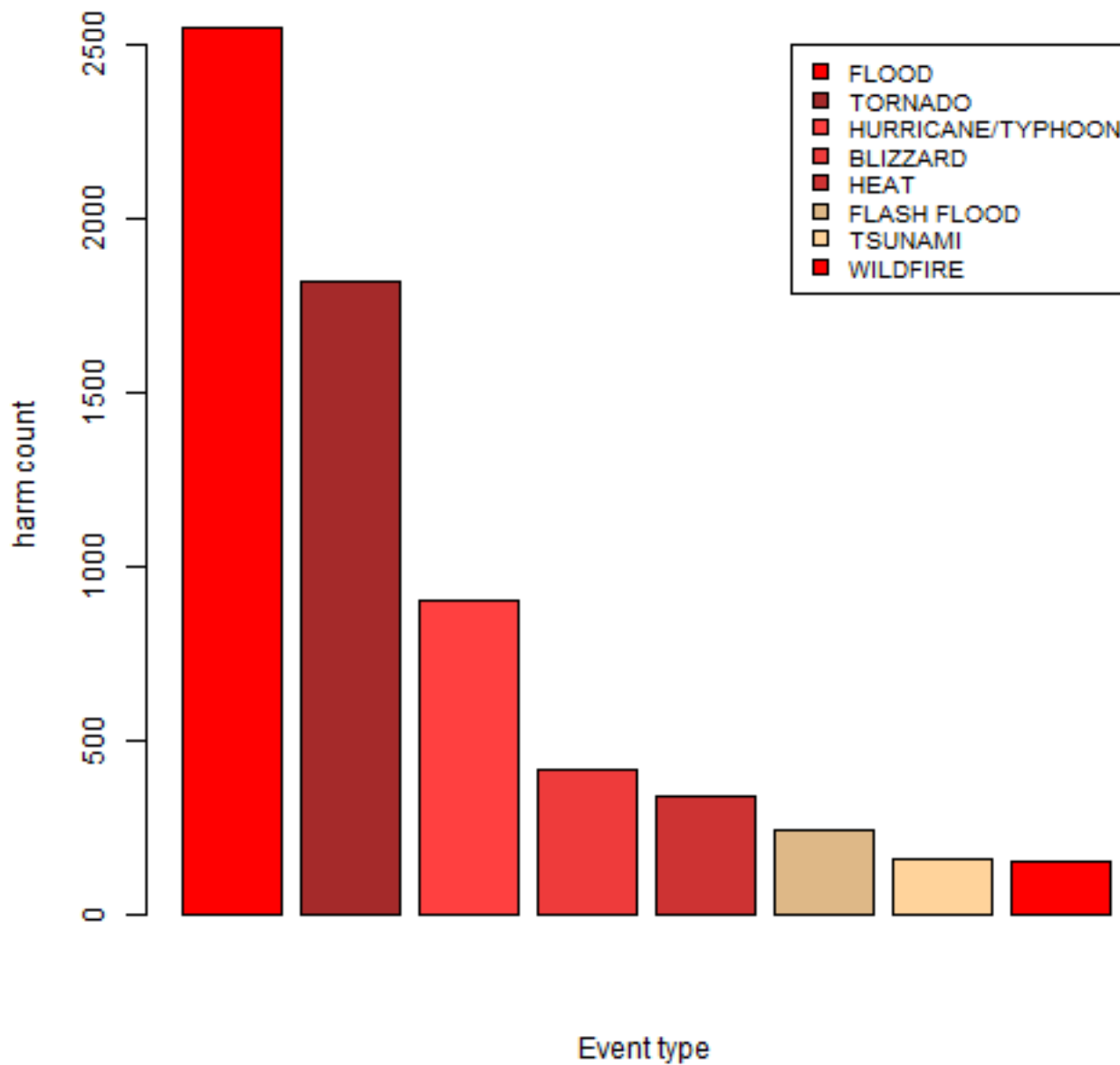
And lets see the overall harm (deaths + injuries)

```
# first lets merge the 2 datasets
```

```
harmPerEvtyp = merge(top10Fatalities,top10Injuries)
harmPerEvtyp$HARM = harmPerEvtyp$FATALITIES + harmPerEvtyp$INJURIES
# order the results
harmPerEvtyp = harmPerEvtyp[with(harmPerEvtyp, order(-HARM)), ]

barplot(harmPerEvtyp$HARM,main="Top-10 harmful events",
        xlab="Event type",ylab="harm count", col=c("red", "brown","brown1","brown2",
"brown3","burlywood","burlywood1"), legend=harmPerEvtyp[, "event"],
        args.legend=c(cex=0.8))
```

Top-10 harmful events



As we can see, overall, the 10-most harmful event is the tornado, following is excessive heat, wind, flood, lightening, heat and flash flood.

**We should take very special care regarding tornados!!!**

## Events with greatest economic consequences

Now lets see what happens with the economic aspect. The question is: **Across the United States, which types of events have the greatest economic consequences?** Let's see for each economic factor in each own:

The fields we are interested in are:

- 1.“PROPDMG” (property damage)
- 2.“CROPDMG” (crop damage)

with corresponding units: “PROPDMGEXP” (unit for property damage) “CROPDMGEXP” (unit for crop damage)

So this is the property damage per event type:

```
propDmgPerEvtyp = aggregate(weather_dataset[, 'PROPDMG'],
  by=list(event = weather_dataset$EVTYPE), FUN=sum)

# order the results
propDmgPerEvtyp = propDmgPerEvtyp[with(propDmgPerEvtyp, order(-PROPDMG)), ]
top10propDmg = head(propDmgPerEvtyp, n=10)
top10propDmg
```

```
##           event PROPDMG
## 15      TORNADO 1051902
## 8       HIGH WIND  948690
## 16    TROPICAL STORM  628520
## 4        FLOOD   221000
## 10     HURRICANE  140250
## 20     WILDFIRE  125121
## 3      FLASH FLOOD   97712
## 19      TSUNAMI   81000
## 14 THUNDERSTORM WINDS  75680
## 22 WINTER STORM HIGH WINDS  60000
```

As we can see, the tornado and flash wind are the events with the greatest damage

Lets see what about the crop damage

```
cropDmgPerEvttyp = aggregate(weather_dataset[ 'CROPDMG' ],
  by=list(event = weather_dataset$EVTYPE), FUN=sum)

# order the results
cropDmgPerEvttyp = cropDmgPerEvttyp[with(cropDmgPerEvttyp, order(-CROPDMG)), ]

top10cropDmg = head(cropDmgPerEvttyp, n=10)
top10cropDmg
```

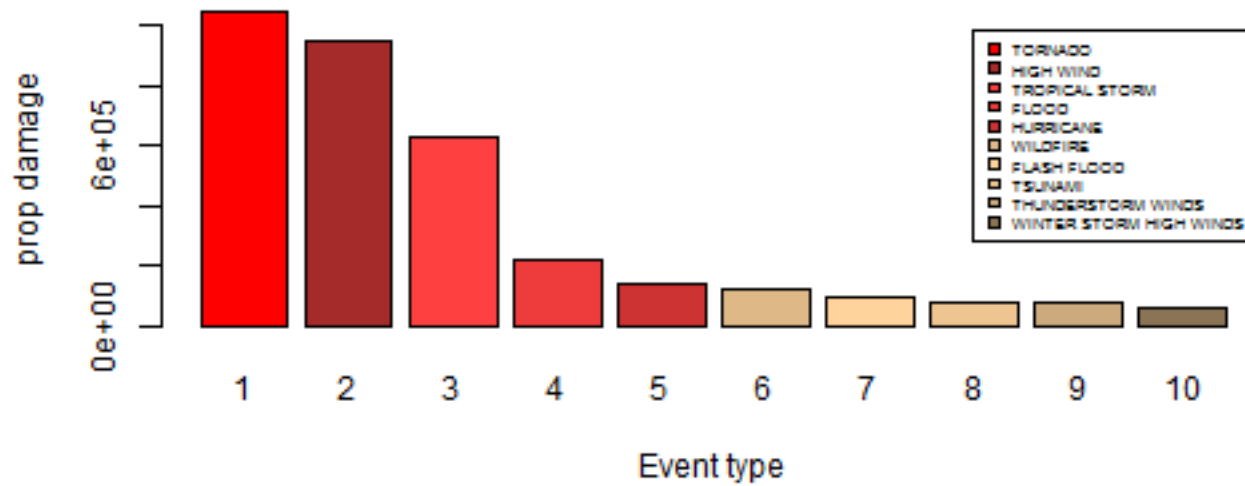
```
##           event CROPDMG
## 2    EXCESSIVE HEAT 492400
## 11  HURRICANE/TYPHOON 285002
## 8      HIGH WIND 222935
## 10     HURRICANE 127000
## 16    TROPICAL STORM 121695
## 1      BLIZZARD 105000
## 15     TORNADO 93525
## 20     WILDFIRE 75150
## 14 THUNDERSTORM WINDS 50016
## 4           FLOOD 17731
```

And we can see that Hail is the event with the greatest damage, following is flsh flood, wind, tornado, etc.

Let see a plot of all this:

```
par(mfrow = c(2, 1))
# prop plot
barplot(top10propDmg$PROPDMG,main="Top-10 prop damage per event type",
  xlab="Event type", ylab="prop damage",col=c("red", "brown", "brown1", "brown2",
"brown3", "burlywood", "burlywood1", "burlywood2", "burlywood3", "burlywood4"), names.arg=1:10,
legend=top10propDmg[, "event"], args.legend=c(cex=0.5))
# and this is the crop plot
barplot(top10cropDmg$CROPDMG,main="Top-10 Crop damage per event type",
  xlab="Event type", ylab="crop damage",col=c("red", "brown", "brown1", "brown2",
"brown3", "burlywood", "burlywood1", "burlywood2", "burlywood3", "burlywood4"), names.arg=1:10,
legend=top10cropDmg[, "event"], args.legend=c(cex=0.5))
```

### Top-10 prop damage per event type



### Top-10 Crop damage per event type

