Storm Data Project

Francesca Chiappetta

1/30/2021

Tornados are the most harmful severe weather events to human health & floods cause the most economic damage

Synopsis Tornados cause the highest number of deaths and the highest number of injuries among all severe weather types since 1950 to November 2011. The severe weather events that cause the highest number of fatalities are tornados, excessive heat, flash floods, heat, lightening, TSTM wind, floods, rip currents, high winds, and avalanches. The severe weather events that cause the highest number of injuries are tornados, TSTM wind, floods, excessive heat, lightening, heat, ice storms, flash floods, thunderstorm winds, and hail.

Floods have caused the highest economic damange (measured by property and crop damage) among all severe weather types since 1950 to November 2011. The severe weather events that cause the highest economic damage (measured by property and crop damage) are floods, hurricaines/typhoons, tornados, storm surges, hail, flash floods, droughts, hurricaines, river floods, and ice storms.

```
#1 Create a directory called data if it doesnt exist
#2 Save URL to variable fileUrl
#3 download file into data directory
#4 use fread to read in data.table package (fread can import bz2 files directly)
library(data.table) #import dat.table library

if(!file.exists("data")){dir.create("data")} #1
fileUrl <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2" #2
download.file(fileUrl, destfile = "./data/stormdata.csv.bz2", mode = "wb") #3
stormdata <- fread(file = "./data/stormdata.csv.bz2", header = TRUE, stringsAsFactors = FALSE, sep = ",
```

Data Processing

```
colnames(stormdata) #check names of columns for dataset
```

Data Analysis

```
##  [1] "STATE__"    "BGN_DATE"   "BGN_TIME"   "TIME_ZONE"  "COUNTY"
##  [6] "COUNTYNAME" "STATE"      "EVTYPE"     "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"
```

```
#subset my data for columns im going to use in my analysis
subset_stormdata <- stormdata[, c("EVTYPE", "FATALITIES", "INJURIES", "PROPDMG", "PROPDMGEXP", "CROPDMG

#Check number of rows and if any columnshave NA values and how many
nrow(subset_stormdata) #902,297 rows
```

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

```
sapply(subset_stormdata, function(x) sum(is.na(x))) #there are some NA values in PROPDMGEXP AND CROPDMG
```

```
##     EVTYPE FATALITIES   INJURIES    PROPDMG PROPDMGEXP    CROPDMG CROPDMGEXP
##          0          0          0          0          0          0          0
```

```
#examine the data
str(subset_stormdata) #will give us class
```

```
## Classes 'data.table' and 'data.frame':   902297 obs. of  7 variables:
##  $ EVTYPE    : chr  "TORNADO" "TORNADO" "TORNADO" "TORNADO" ...
##  $ 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  "" "" "" "" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
summary(subset_stormdata) #give some insight into data trends
```

```
##     EVTYPE            FATALITIES          INJURIES            PROPDMG
##  Length:902297      Min.   :  0.0000   Min.   :   0.0000   Min.   :   0.00
##  Class :character   1st Qu.:  0.0000   1st Qu.:   0.0000   1st Qu.:   0.00
##  Mode  :character   Median :  0.0000   Median :   0.0000   Median :   0.00
##                     Mean   :  0.0168   Mean   :   0.1557   Mean   :  12.06
##                     3rd Qu.:  0.0000   3rd Qu.:   0.0000   3rd Qu.:   0.50
##                     Max.   :583.0000   Max.   :1700.0000   Max.   :5000.00
##   PROPDMGEXP           CROPDMG         CROPDMGEXP
##  Length:902297      Min.   :  0.000   Length:902297
##  Class :character   1st Qu.:  0.000   Class :character
##  Mode  :character   Median :  0.000   Mode  :character
##                     Mean   :  1.527
##                     3rd Qu.:  0.000
##                     Max.   :990.000
```

```
#check the unique values of PROPDMGEXP and CROPDMGEXP
unique(subset_stormdata$PROPDMGEXP)
```

How To Handle Values of PROPDMGEXP and CROPDMGEXP

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

```
unique(subset_stormdata$CROPDMGEXP)
```

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

- After doing some research, reading the NOAA storm data documentation, and direction from the RPubs publication *How To Handle Exponent Value of PROPDMGEXP and CROPDMGEXP*, I have determined that the coded values in the PROPDMGEXP AND CROPDMGEXP values are multipliers for the data
- *Note: EXP = exponent*
- *Note: any numeric value that is not 1 and 0 (2;3;4;5;6;7;8) is believed to be improper handling in the dataset that was later fixed in a 2012 update. The error is that those numbers were intended to be the ones digit in the corresponding PROPDMG and CROPDMG values. This would make each PROPDMG and CROPDMG value with a corresponding CROPDMGEXP and PROPDMGEXP value to be approximately 10x the reported value. Therefore, those numbers will be a multiplier of ten*

These are possible values of CROPDMGEXP and PROPDMGEXP: H,h,K,k,M,m,B,b,+,-,?,0,1,2,3,4,5,6,7,8, and blank-character; H,h = hundreds = 100; K,k = kilos = thousands = 1,000; M,m = millions = 1,000,000; B,b = billions = 1,000,000,000; (+) = 1; (-) = 0; (?) = 0; black/empty character = 0; 2,3,4,5,6,7,8 = 10

```
#convert PROPDMGEXP and CROPDMGEXP values and multiply PROPDMG by PROPDMGEXP and amd CROPDMG by CROPDMG
subset_stormdata$PROPDMGEXP[is.na(subset_stormdata$PROPDMGEXP)] <- 0
subset_stormdata$PROPDMGEXP[subset_stormdata$PROPDMGEXP == ""] <- 1
subset_stormdata$PROPDMGEXP[grep("[-+?]", subset_stormdata$PROPDMGEXP)] <- 1
subset_stormdata$PROPDMGEXP[grep("[2-8]", subset_stormdata$PROPDMGEXP)] <- 10
subset_stormdata$PROPDMGEXP[grep("[Hh]", subset_stormdata$PROPDMGEXP)] <- 100
subset_stormdata$PROPDMGEXP[grep("[Kk]", subset_stormdata$PROPDMGEXP)] <- 1000
subset_stormdata$PROPDMGEXP[grep("[Mm]", subset_stormdata$PROPDMGEXP)] <- 1000000
subset_stormdata$PROPDMGEXP[grep("[Bb]", subset_stormdata$PROPDMGEXP)] <- 1000000000

subset_stormdata$CROPDMGEXP[is.na(subset_stormdata$CROPDMGEXP)] <- 0
subset_stormdata$CROPDMGEXP[subset_stormdata$CROPDMGEXP == ""] <- 1
subset_stormdata$CROPDMGEXP[grep("[?]", subset_stormdata$CROPDMGEXP)] <- 1
subset_stormdata$CROPDMGEXP[grep("[2]", subset_stormdata$CROPDMGEXP)] <- 10
subset_stormdata$CROPDMGEXP[grep("[Hh]", subset_stormdata$CROPDMGEXP)] <- 100
subset_stormdata$CROPDMGEXP[grep("[Kk]", subset_stormdata$CROPDMGEXP)] <- 1000
subset_stormdata$CROPDMGEXP[grep("[Mm]", subset_stormdata$CROPDMGEXP)] <- 1000000
subset_stormdata$CROPDMGEXP[grep("[Bb]", subset_stormdata$CROPDMGEXP)] <- 1000000000

#check unique values again
unique(subset_stormdata$PROPDMGEXP)
```

```
## [1] "1000"  "1e+06" "1"     "1e+09" "0"     "10"    "100"
```

```
unique(subset_stormdata$CROPDMGEXP)
```

```
## [1] "1"     "1e+06" "1000"  "1e+09" "0"     "10"
```

```
#create new columns for property damage and crop damage and change PROPDMGEXP and CROPDMGEXP to numeric
subset_stormdata$PROPDMGEXP <-as.numeric(subset_stormdata$PROPDMGEXP)
subset_stormdata$Property.Damage <- subset_stormdata$PROPDMG * subset_stormdata$PROPDMGEXP
subset_stormdata$CROPDMGEXP <-as.numeric(subset_stormdata$CROPDMGEXP)
subset_stormdata$Crop.Damage <- subset_stormdata$CROPDMG * subset_stormdata$CROPDMGEXP

#combine property and crop damage to be a columnn containing total damage
subset_stormdata$Total.Damage <- subset_stormdata$Crop.Damage + subset_stormdata$Property.Damage
```

```
#I want to subset my data and change column names for sake of clarity
stormdata_final <- subset_stormdata[, c("EVTYPE", "FATALITIES", "INJURIES", "Total.Damage")]
colnames(stormdata_final) <- c("Event.Type", "Fatalities", "Injuries", "Total.Damage")
```

```
sum_fatalities <- aggregate(Fatalities ~ Event.Type, stormdata_final, sum)
sum_fatalities <- sum_fatalities[order(-sum_fatalities$Fatalities),c(1,2)]

sum_injuries <- aggregate(Injuries ~ Event.Type, stormdata_final, sum)
sum_injuries <- sum_injuries[order(-sum_injuries$Injuries),c(1,2)]

sum_totaldamage <- aggregate(Total.Damage ~ Event.Type, stormdata_final, sum)
sum_totaldamage <-sum_totaldamage[order(-sum_totaldamage$Total.Damage),c(1,2)]
```

Calculate the sum of each column and order it in descending order

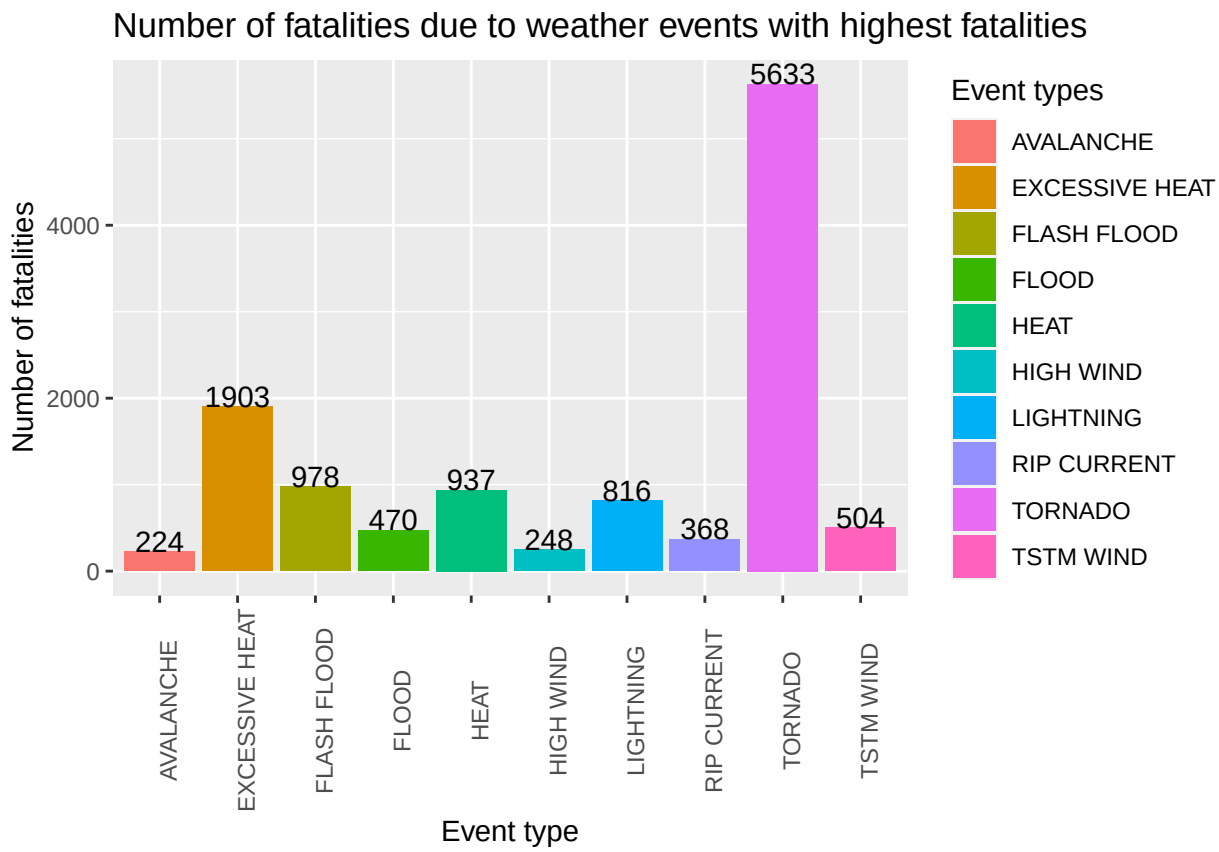
```
#subset top 10
top10_fatalities <- sum_fatalities[1:10,]
top10_injuries <- sum_injuries[1:10,]
top10_damage <- sum_totaldamage[1:10,]
```

Get top 10 most harmful in terms of fatalities, injuries, and economic damage

Results

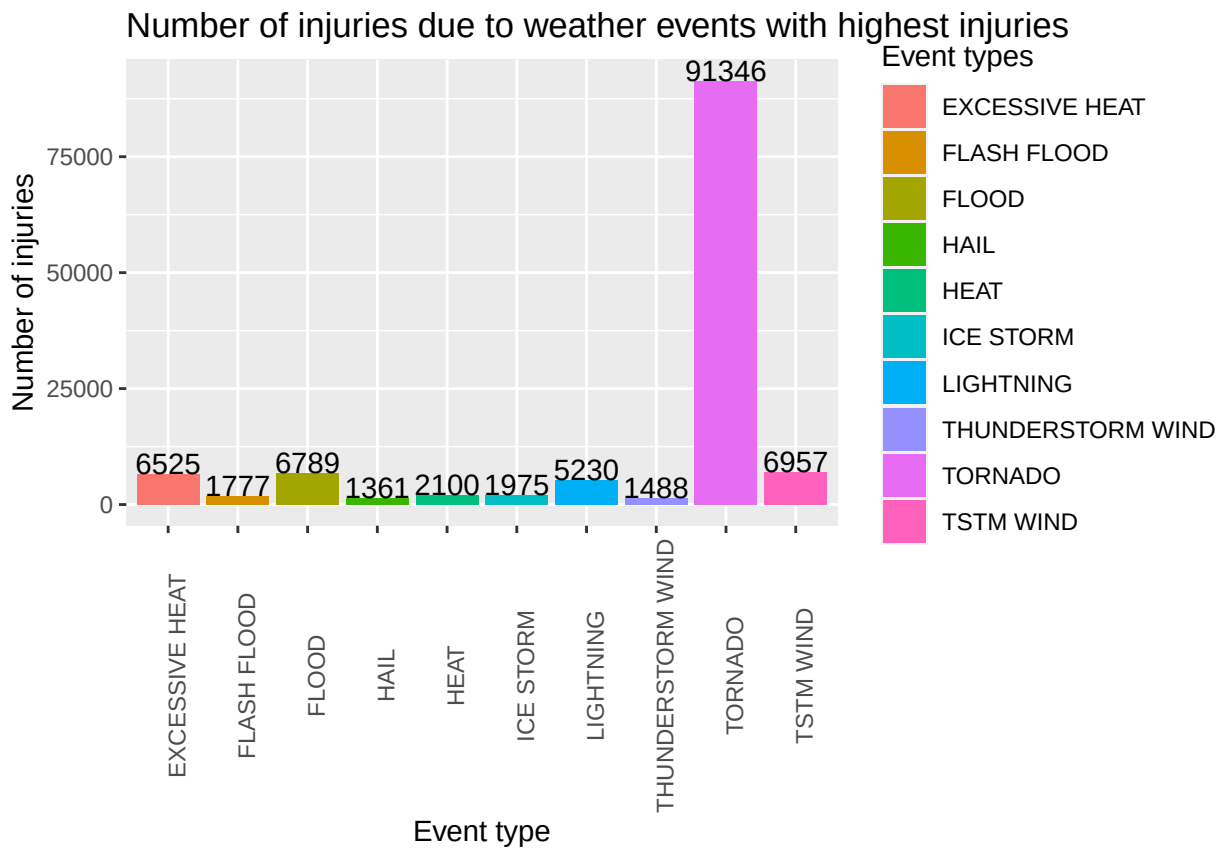
1. Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health?

```
library(ggplot2)

ggplot(top10_fatalities, aes(x = Event.Type, y = Fatalities, fill = Event.Type, label =Fatalities)) +
geom_bar(stat = "identity") +
labs(x = "Event type", y = "Number of fatalities", fill = "Event types") +
ggtitle("Number of fatalities due to weather events with highest fatalities") +
geom_text(aes(label = Fatalities),vjust=0) +
theme(axis.text.x = element_text(angle = 90))
```



Tornados cause the highest number of fatalities among all severe weather types since 1950 to November 2011.

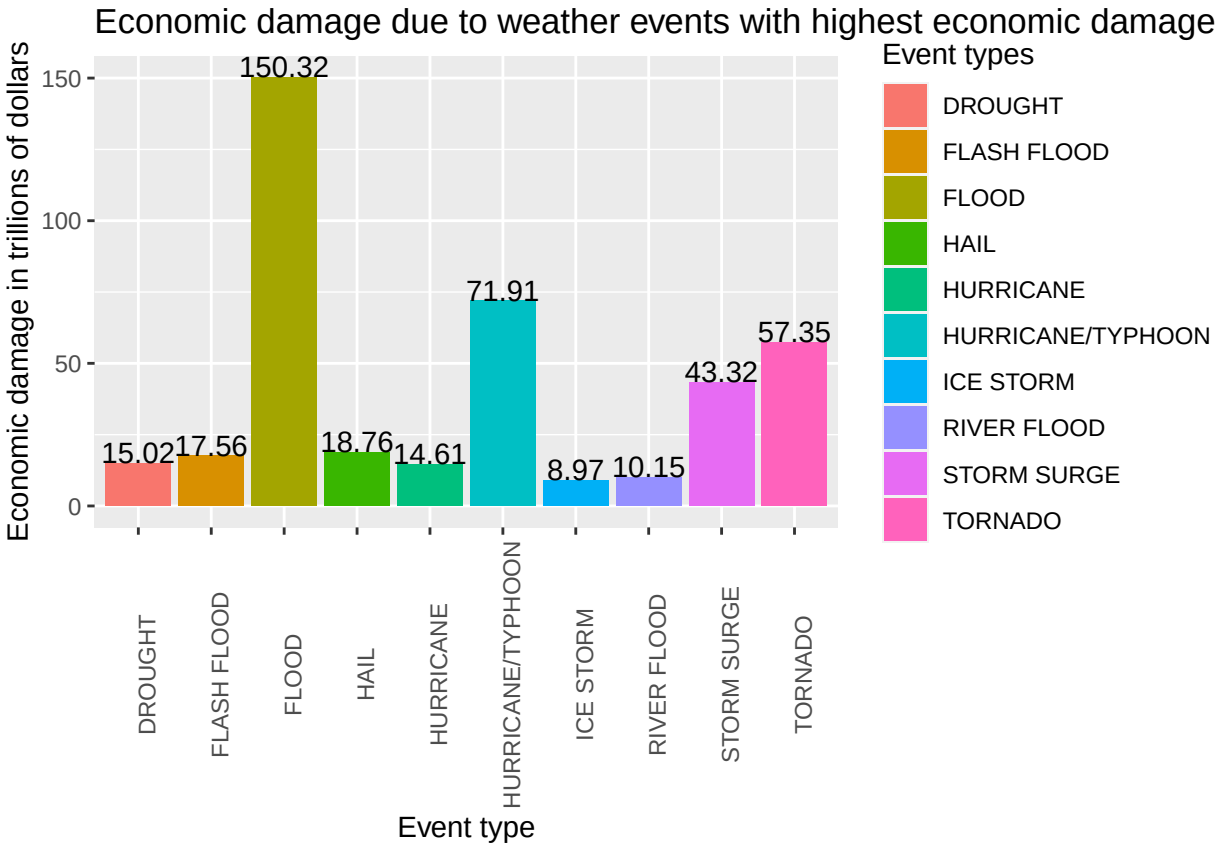
```
library(ggplot2)

ggplot(top10_injuries, aes(x = Event.Type, y = Injuries, fill = Event.Type, label =Injuries)) +
geom_bar(stat = "identity") +
labs(x = "Event type", y = "Number of injuries", fill = "Event types") +
ggtitle("Number of injuries due to weather events with highest injuries") +
geom_text(aes(label = Injuries),vjust=0) +
theme(axis.text.x = element_text(angle = 90))
```



Tornados cause the highest number of injuries among all severe weather types since 1950 to November 2011.

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

```
library(ggplot2)

ggplot(top10_damage, aes(x = Event.Type, y = Total.Damage/1000000000, fill = Event.Type, label =Total.Da
geom_bar(stat = "identity") +
labs(x = "Event type", y = "Economic damage in trillions of dollars", fill = "Event types") +
ggtitle("Economic damage due to weather events with highest economic damage") +
geom_text(aes(label = round(Total.Damage/1000000000,2)),vjust=0) +
theme(axis.text.x = element_text(angle = 90))
```



Floods have caused the highest economic damange (property and crop damage) among all severe weather types since 1950 to November 2011.