Severe Weather Events in the NOAA Storm Database

Andrea Villaroman

4/24/2020

Synopsis

This is report is written for the Reproducible Research class in the Coursera Data Science Specialization by John Hopkins University. The instructions for this project are found on the course website. In this report, we explore the National Oceanic and Atmospheric Administration (NOAA) Storm Database and present our findings in the results below.
As part of the data processing, I have filtered and transformed the original Storm Data for this project. For this analysis, I was most interested in population harm (as represented by injury and fatility) and economic consequence (as represented by crop damage and property damage). Since there were many unique event types with varying features and severity, I attempt to represent the events as simply as possible while still retaining useful and actionable information that can be graphically represented.
In a provided graphical representation, I show the top 10 severe weather events by (a) population harm and (b) damages. The most harmful (population-wise and economically) severe weather events were tornadoes, thunderstorms, and flooding. Tornadoes were by far the most harmful in terms of fatalities and injuries, followed by heat.

The Question(s)

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

To answer this question, we consider the given data from injuries and fatalities. Data is quantified by a count of the number of injuries or fatalities.

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

To answer this question, we consider the given data from crop damage and property damage. Data is quantified in dollar amounts, according to Storm Data Documentation with a separate column listing the magnitude (K = thousands, M = millions, B = billions).

Data Processing

Data Source and Loading Original Data

The data was unzipped and loaded into R using read.csv. The Storm Data Documentation provides a comprehensive description of the available data.

```
zipfile <- "repdata_data_StormData.csv.bz2"

repdata_stormdata <- read.csv(zipfile)
str(repdata_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 1
##  $ 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 8
##  $ 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 ...
```

Filtering and Transforming the Data

The data processing consisted of the following methodology:
1. Filtering the data to select for complete (more recent) events, population harm (fatalities and injuries), and damage (crop damage and property damage).
2. Converting property damage and crop damage values to a common dollar magnitude.
3. Identifying and grouping EVTYPE factors.
4. Summarizing by event type: (a) population harm and (b) economic consequence.

1. Filtering the Original Dataset

Although the dataset contains data from 1950 up until November 2011, we want to consider the most complete set of data. Below a histogram of datapoints distribution per year:

```
# get data we want

years <-
        repdata_stormdata %>%
        mutate(year = as.numeric(format(as.Date(BGN_DATE, '%m/%d/%Y'),'%Y'))) %>%
        select(REFNUM, year)
hist(years$year
     , xlab = "Year"
     , ylab = "Datapoints"
     , main = "Histogram of Storm Datapoints per Year"
)

dist <- quantile(years$year, probs = c(0.1,0.25,0.5,0.75))
abline(v=dist[["10%"]], col = 'darkslategray1', lty=2, lwd=3)
abline(v=dist[["25%"]], col = 'darkslategray2', lty=2, lwd=3)
abline(v=dist[["50%"]], col = 'darkslategray3', lty=2, lwd=3)
abline(v=dist[["75%"]], col = 'darkslategray4', lty=2, lwd=3)
legend(1950, 200000
       , legend=c("10% Quantile", "25% Quantile", "50% Quantile", "75% Quantile")
       , col=c("darkslategray1", "darkslategray2", "darkslategray3", "darkslategray4")
       , lty=2
       , lwd=3)
```

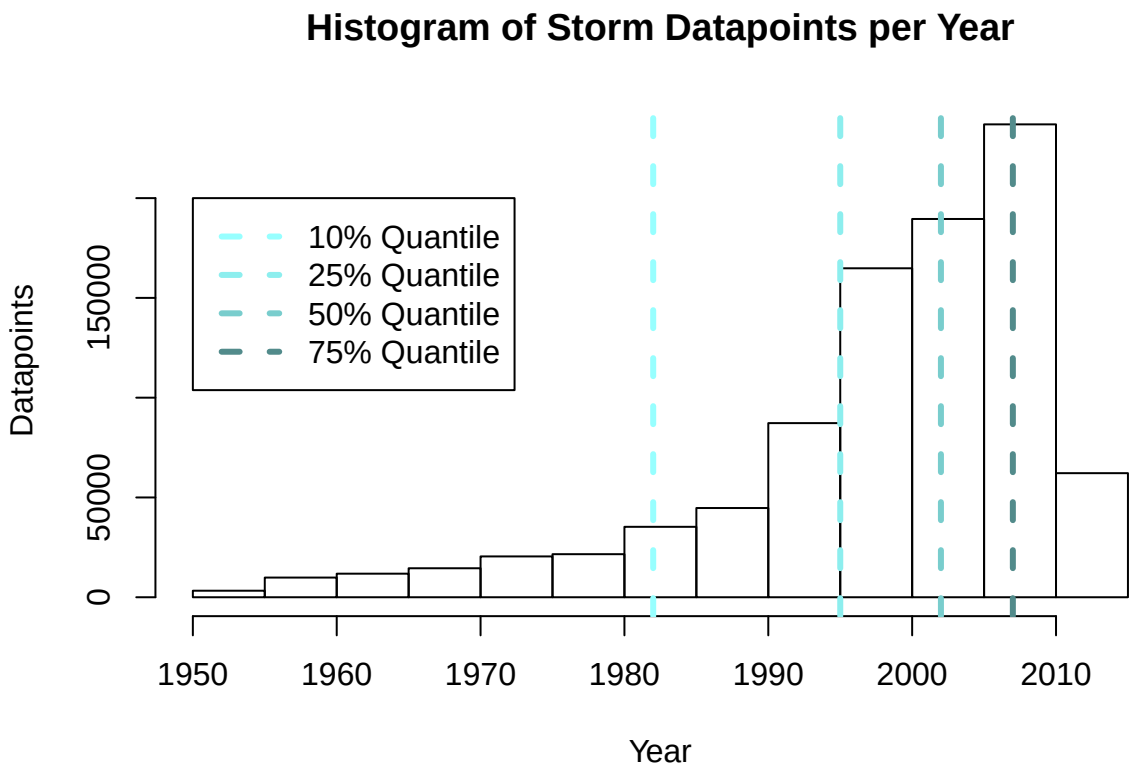

Figure 1. Histogram of Storm Datapoints Per Year shows the distribution of datapoints by year and also shows quantile values at 10%, 25%, 50%, and 75%.

Looking at this plot, I choose to ignore data before 1982, which maintains 90% of the dataset and will have a higher likelihood of completeness.

```
quantile <- "10%"
keep <- years %>% filter(year >= dist[[quantile]]) %>% select(REFNUM)

stormdata <-
        repdata_stormdata %>%
        filter(REFNUM %in% keep[["REFNUM"]]) %>%
        select(EVTYPE
               , FATALITIES
               , INJURIES
               , PROPDMG
               , PROPDMGEXP
               , CROPDMG
               , CROPDMGEXP
               , REFNUM)

str(stormdata)
```

```
## 'data.frame':    816268 obs. of  8 variables:
##  $ EVTYPE    : Factor w/ 985 levels "   HIGH SURF ADVISORY",..: 834 834 834 834 834 834 244 856 856 8
##  $ FATALITIES: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ INJURIES  : num  0 0 6 0 4 6 0 0 0 0 ...
```

```
##  $ PROPDMG   : num  250 250 250 2.5 25 250 0 0 0 250 ...
##  $ PROPDMGEXP: Factor w/ 19 levels "","-","?","+",..: 17 17 17 17 17 17 1 1 1 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 ...
##  $ REFNUM    : num  2265 2266 2267 2268 2269 ...
```

```
head(stormdata)
```

```
##    EVTYPE FATALITIES INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP REFNUM
## 1 TORNADO          0        0   250.0          K       0                2265
## 2 TORNADO          0        0   250.0          K       0                2266
## 3 TORNADO          0        6   250.0          K       0                2267
## 4 TORNADO          0        0     2.5          K       0                2268
## 5 TORNADO          0        4    25.0          K       0                2269
## 6 TORNADO          0        6   250.0          K       0                2270
```

2. Convert Damages to Common Dollar Magnitude

The conversion values used in this analysis are based this analysis, which also has significant contribution by David Hood and Eddie Song, as cited.

```
# create exp conversion function

convertwithexp <- function(dmg,exp) {
        if (exp == "[hH]") {return(dmg*10^2)}
        if (exp == "[kK]") {return(dmg*10^3)}
        if (exp == "[mM]") {return(dmg*10^6)}
        else return(dmg) # +if (exp == "[bB]") {return(dmg*10^9)}
        if (exp %in% as.character(c(0:8))) {return(dmg*10^1)}
        if (exp %in% c("-", "?", "")) {return(dmg*0)}
}

# add computed damage values

stormdata$propertydamage <- mapply(convertwithexp, stormdata$PROPDMG, stormdata$PROPDMGEXP)
stormdata$cropdamage <- mapply(convertwithexp, stormdata$CROPDMG, stormdata$CROPDMGEXP)
```

3. Further Filtering and Grouping Data

```
# only regard rows that have damage or injury/fatality information
stormdata_filtered <-
        filter(stormdata
               , (propertydamage != 0 | cropdamage !=0) | (FATALITIES !=0 | INJURIES !=0))
```

This helps reduce our dataset to 236476 (down from 816268).

Because of the nature of the dataset, there are several typos in the event labels as well as similar-type events. One such example:

```
head(grep("THUN.*M",unique(repdata_stormdata$EVTYPE), value=TRUE), n=10)
```

```
##  [1] "THUNDERSTORM WINDS"             "THUNDERSTORM WIND"
##  [3] "THUNDERSTORM WINS"              "THUNDERSTORM WINDS LIGHTNING"
##  [5] "THUNDERSTORM WINDS/HAIL"        "THUNDERSTORM WINDS HAIL"
##  [7] "FLASH FLOODING/THUNDERSTORM WI" "THUNDERSTORM"
##  [9] "THUNDERSTORM WINDS/FUNNEL CLOU" "SEVERE THUNDERSTORM"
```

We therefore use our best judgement in grouping event types using a custom function. Below, I show the function for grouping events and print all the unique groups after applying the function:

```
# group event types function

type_event <- function(x) {

        evtype <- as.character(x)

        if (is.null(evtype) | evtype == "?") {return("OTHER")}
        if (grepl("TORN.*O", evtype, ignore.case=TRUE)) {return("TORNADO")}
        if (grepl("TSTM|THUN.*M|MICROBURST", evtype, ignore.case=TRUE))
                {return("THUNDERSTORM AND/OR TSTM WINDS")}
        if (grepl("HEAT|WARM", evtype, ignore.case=TRUE)) {return("HEAT")}
        if (grepl("AVALAN", evtype, ignore.case=TRUE)) {return("AVALANCHE")}
        if (grepl("HURRICANE", evtype, ignore.case=TRUE)) {return("HURRICANE")}
        if (grepl("RAIN|SHOWER", evtype, ignore.case=TRUE)) {return("RAIN")}
        if (grepl("HAIL", evtype, ignore.case=TRUE)) {return("HAIL")}
        if (grepl("FLOOD", evtype, ignore.case=TRUE)) {return("FLOOD")}
        if (grepl("LI.*ING", evtype, ignore.case=TRUE)) {return("LIGHTNING")}
        if (grepl("FOG", evtype, ignore.case=TRUE)) {return("FOG")}
        if (grepl("SNOW", evtype, ignore.case=TRUE)) {return("SNOW")}
        if (grepl("FROST", evtype, ignore.case=TRUE)) {return("FROST")}
        if (grepl("FLOOD", evtype, ignore.case=TRUE)) {return("FLOOD")}
        if (grepl("FREEZ", evtype, ignore.case=TRUE)) {return("FREEZE")}
        if (grepl("FIRE", evtype, ignore.case=TRUE)) {return("FIRE")}
        if (grepl("DROUGHT|DRY", evtype, ignore.case=TRUE))
                {return("DROUGHT/DRY")}
        if (grepl("COLD|CHILL|WINT|HYPOTHERMIA|LOW TEMP|COOL", evtype, ignore.case=TRUE))
                {return("COLD")}
        if (grepl("WIND", evtype, ignore.case=TRUE)) {return("WIND")}
        if (grepl("ICE|ICY", evtype, ignore.case=TRUE)) {return("ICE")}
        if (grepl("TROPICAL STORM", evtype, ignore.case=TRUE))
                {return("TROPICAL STORM")}
        if (grepl("HIGH TIDE|HIGH SURF|HIGH WATER|SURGE/TIDE|HIGH SEAS"
                  , evtype, ignore.case=TRUE)) {return("HIGH TIDE/SURF")}
        if (grepl("MUDSLIDE|MUD SLIDE", evtype, ignore.case=TRUE)) {return("MUDSLIDES")}
        if (grepl("LANDSLIDE", evtype, ignore.case=TRUE)) {return("LANDSLIDES")}

        evtype
}

# use uppercase
stormdata2 <- as.data.table(stormdata_filtered)
```

```r
# apply function to each row of EVTYPE
stormdata2$eventtype <- mapply(type_event, stormdata2$EVTYPE)

# look at new grouping
unique(stormdata2$eventtype)
```

```
##  [1] "TORNADO"                       "THUNDERSTORM AND/OR TSTM WINDS"
##  [3] "HAIL"                          "FLOOD"
##  [5] "COLD"                          "HURRICANE"
##  [7] "RAIN"                          "LIGHTNING"
##  [9] "FOG"                           "RIP CURRENT"
## [11] "HEAT"                          "WIND"
## [13] "WATERSPOUT"                    "FREEZE"
## [15] "AVALANCHE"                     "MARINE MISHAP"
## [17] "HIGH TIDE/SURF"                "SEVERE TURBULENCE"
## [19] "SNOW"                          "DUST STORM"
## [21] "APACHE COUNTY"                 "SLEET"
## [23] "DUST DEVIL"                    "ICE"
## [25] "FIRE"                          "HIGH"
## [27] "MUDSLIDES"                     "FUNNEL CLOUD"
## [29] "HEAVY SURF"                    "DROUGHT/DRY"
## [31] "BLIZZARD"                      "STORM SURGE"
## [33] "WATERSPOUT-"                   "TROPICAL STORM"
## [35] "FROST"                         "EXCESSIVE WETNESS"
## [37] "GUSTNADO"                      "GROUND BLIZZARD"
## [39] "DUST DEVIL WATERSPOUT"         "GLAZE"
## [41] "RIP CURRENTS/HEAVY SURF"       "URBAN AND SMALL"
## [43] "HEAVY MIX"                     "RIP CURRENTS"
## [45] "COASTAL SURGE"                 "HEAVY PRECIPITATION"
## [47] "HIGH WAVES"                    "RAPIDLY RISING WATER"
## [49] "LANDSLIDES"                    "HEAVY SEAS"
## [51] "OTHER"                         "URBAN/SMALL STREAM"
## [53] "HEAVY SWELLS"                  "URBAN SMALL"
## [55] "Other"                         "URBAN/SML STREAM FLD"
## [57] "ROUGH SURF"                    "Heavy Surf"
## [59] "Dust Devil"                    "Marine Accident"
## [61] "COASTAL STORM"                 "Beach Erosion"
## [63] "Landslump"                     "Coastal Storm"
## [65] "Glaze"                         "MIXED PRECIP"
## [67] "DOWNBURST"                     "Mixed Precipitation"
## [69] "COASTALSTORM"                  "DAM BREAK"
## [71] "TYPHOON"                       "HIGH SWELLS"
## [73] "COASTAL EROSION"               "SEICHE"
## [75] "HYPERTHERMIA/EXPOSURE"         "ROCK SLIDE"
## [77] "LANDSPOUT"                     "MIXED PRECIPITATION"
## [79] "ROUGH SEAS"                    "ROGUE WAVE"
## [81] "BLOWING DUST"                  "VOLCANIC ASH"
## [83] "HAZARDOUS SURF"                "DROWNING"
## [85] "TROPICAL DEPRESSION"           "TSUNAMI"
## [87] "ASTRONOMICAL LOW TIDE"         "DENSE SMOKE"
```

4. Summarizing by Event Type

Using this new dataset, we can now summarize the fatality and injury information by event type:

```
# summarize fatalities and injuries by the event type

populationharm <-
        stormdata2  %>%
        select(eventtype, INJURIES, FATALITIES) %>%
        mutate(harmed = INJURIES + FATALITIES) %>%
        group_by(eventtype) %>%
        summarise_all(sum) %>%
        arrange(desc(harmed))
```

I also summarize the crop damage and property damage by event type:

```
# summarize property damage and crop damage

econdamage <-
        stormdata2 %>%
        select(eventtype, propertydamage, cropdamage) %>%
        mutate(totaldamage = propertydamage + cropdamage) %>%
        group_by(eventtype) %>%
        summarise_all(sum) %>%
        arrange(desc(totaldamage))
```

Results

Population Harm

After summarizing the results, I can arrange and get the top 10 severe weather events for each event type and graphically represent it in stacked bar plots:

```
top10_harmed <- reshape2::melt(head(populationharm, n=10)
                               , measure=c("INJURIES", "FATALITIES")
                               , id= "eventtype")

ggplot(top10_harmed, aes(x= reorder(eventtype, value), y= value/1000, fill=variable )) +
        geom_bar(position="stack", stat="identity") +
        labs(title = "Top 10 Severe Weather Events with Highest Population Harm"
        , subtitle= "1982-2011"
        , x = "Event Type"
        , y = "Count(Thousands)") +
        theme(plot.title = element_text(size=12, hjust=0.5)
              , legend.position="top"
              , plot.subtitle = element_text(hjust=0.5)) +
        scale_fill_discrete(name="") +
        coord_flip()
```

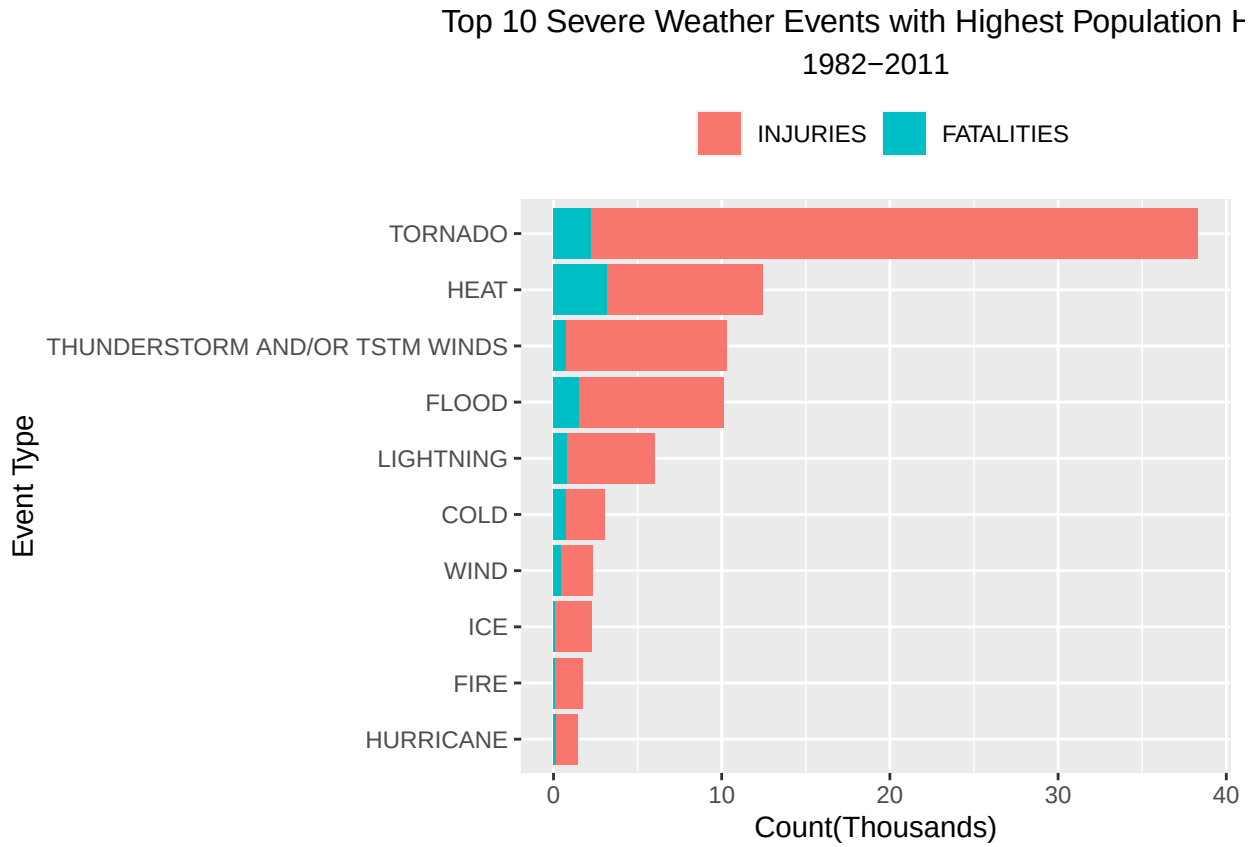

Figure 2. Top 10 Severe Weather Events With Highest Population Harm gives the count of fatality and injury reports over 1982-2011.

Returning to Question 1:

1. Filtering the data to select for event type, population harm (fatalities and injuries), and damage (crop damage and property damage).

Tornadoes appear to cause the greatest amount of population harm, with the vast majority of harm being from reported injuries at nearly 40,000 from 1982-2011. Heat causes the greatest amount of fatality and is also the second highest event causing population harm at ~12,000 from 1982-2011. Other significant events are thunderstorms, floods, lightning, cold, wind, ice, fire, and hurricanes.

```
top10_dmg <- reshape2::melt(head(econdamage, n=10)
                            , measure=c("propertydamage", "cropdamage")
                            , id= "eventtype")

ggplot(top10_dmg, aes(x= reorder(eventtype, value), y= value/(10^6), fill=variable )) +
        geom_bar(position="stack", stat="identity") +
        labs(title = "Top 10 Severe Weather Events With Largest Economic Consequence"
        , subtitle= "1982-2011"
        , x = "Event Type"
        , y = "Damage (Millions of Dollars)") +
        theme(plot.title = element_text(size=12, hjust=0.5)
              , legend.position="top"
              , plot.subtitle = element_text(hjust=0.5)) +
```

```
        scale_fill_discrete(name="",labels = c("Property Damage", "Crop Damage")) +
        coord_flip()
```

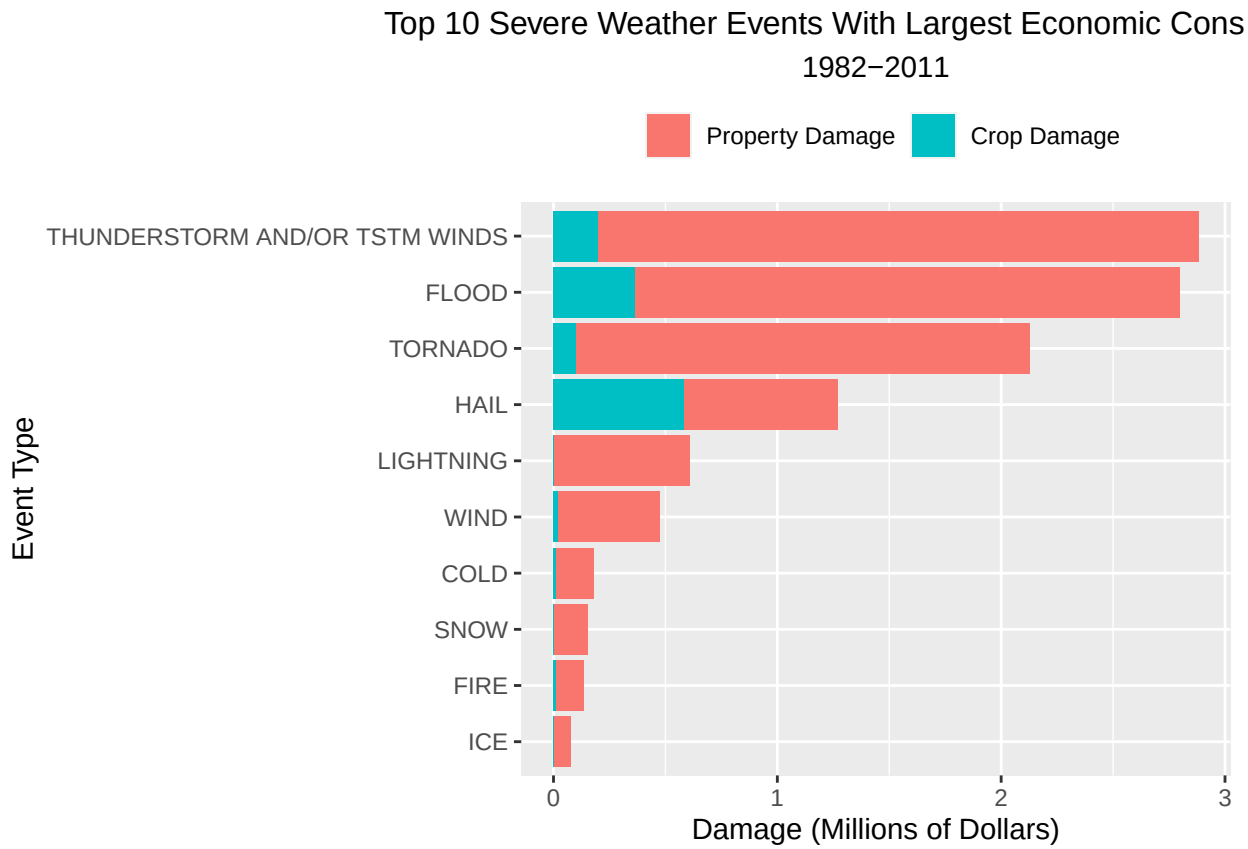

Figure 3. Top 10 Severe Weather Events With Largest Economic Consequences show total property or crop damage values in millions of dollars (1982-2011).

Returning to Question 2:

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

Thunderstorms and respective weather events (such as winds) have caude the greatest economic consequences from 1982-2011 at nearly 3 million dollars. Following closely are floods and tornadoes. Hail, lightning, other wind events, cold, snow, fire, and ice follow respectively in damage. Interestingly, hail caused a significant amount of crop damage (the highest), followed by flooding.
