Data Discovery and Preparation for NOAA Storm Impact Visualization Application analyzing US Storms based on Population Health and Economic Consequences

Mike Pennell

July 3, 2016

Summary

This process performs data discovery and preparation for the NOAA Storm Impact Visualization app. It assesses which types of events were most harmful with respect to population health and economic consequences. The visualization is based on the total economic damages ($), number of people harmed (injured or killed) and number of fatalities.

Data and Documentation

The National Weather Service Storm Data Documentation provides additional information on the source dataset:

documentation: https://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2Fpd01016005curr.pdf
data source: https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2

Data Processing

The data was downloaded from the website above. The property damage (PROPDMG) and crop damage (CROPDMG) were converted to dollar values. Outliers were assessed for impact and one outlier (2006 Napa Flood) was removed due to obvious data entry error.

For each event type (EVTYPE) the following measures were calculated:
* Damages ($): includes damage to crops and property
* Harmed: includes total injured and fatalities
* Fatalities: includes only fatalities
* Count - total number of events of each type

The following aggregations were calculated:
* Total: sum for all events of this type
* Mean Overall: mean for all events of this type
* Mean Top Decile: to assess if significant occurances can be particularly dangerous, the mean of the top decile is calculated
* Max: to evaluate the most impact of any particular event of that type

```
# Initialization
suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(ggplot2))
suppressPackageStartupMessages(library(gridExtra))
```

```
# remove following line before publising
setwd("~/OneDrive/Documents/0 SourceThought Private/Data Science Course/Data Products/Storm App")

highlights <- c("EVTYPE", "BGN_DATE", "STATE", "COUNTYNAME", "BGN_LOCATI", "FATALITIES", "INJURIES", "t
```

```
# Download power storm data from website
fileURL <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
if(!file.exists("dataset.csv.bz2")) {
  download.file(fileURL, destfile = "dataset.csv.bz2", method="curl") #For mac add method="curl"
}
storm <- read.csv("dataset.csv.bz2", stringsAsFactors = FALSE)

# Convert property damage (PROPDMG) and crop damage (CROPDMG) to the real dollar values
# The conversion used the corresponding code (PROPDMGEXP and CROPDMGEXP) to multipy the value entered.
conversion <- data.frame(
   exp = c("", "-", "?", "+", "0", "1", "2", "3", "4", "5", "6", "7", "8", "h", "H", "k",  "K",  "m",
   mul = c(0,  0,   0,   1,   10,  10,  10,  10,  10,  10,  10,  10,  10,  100, 100, 1000, 1000, 100000

storm2 <- merge(storm, conversion, by.x = "PROPDMGEXP", by.y = "exp")
names(storm2)[names(storm2) == "mul"] <- "propMul"
storm2$propDmgAmt = storm2$PROPDMG * storm2$propMul
storm2 <- merge(storm2, conversion, by.x = c("CROPDMGEXP"), by.y = c("exp"))
names(storm2)[names(storm2) == "mul"] <- "cropMul"
storm2$cropDmgAmt = storm2$CROPDMG * storm2$cropMul

# Calculate the total damages (propery + crop) and total people harmed (fatalities + injured)
storm2$totalDamage = storm2$propDmgAmt + storm2$cropDmgAmt
storm2$totalHarmed = storm2$FATALITIES + storm2$INJURIES
```

Outlier Removal

The results are skewed by an incorrect entry. The 2006 Napa flood (see printout) was clearly misslabled as B for billions which significantly impacts calculations. This outlier was removed.

```
#Outlier removal
outlier <- which(storm2$totalDamage == max(storm2$totalDamage))
outlierValue <-storm2[outlier,]
storm2 <- storm2[-outlier,]
print(t(outlierValue[append(highlights,c('PROPDMG', 'PROPDMGEXP'))]))
```

```
##             900796
## EVTYPE      "FLOOD"
## BGN_DATE    "1/1/2006 0:00:00"
## STATE       "CA"
## COUNTYNAME  "NAPA"
## BGN_LOCATI  "COUNTYWIDE"
## FATALITIES  "0"
## INJURIES    "0"
## totalDamage "115032500000"
## REMARKS     "Major flooding continued into the early hours of January 1st, before the Napa River fina
## PROPDMG     "115"
## PROPDMGEXP  "B"
```

Review remaining potential outliers

The next highest values for damages and people harmed (see printouts) appear reasonable based on review and so should not skew results.

```
# Next most harmful events in terms of damage and people harmed are realistic (see printout).
# Significant outliers appear to have been removed.

# Event with maximum damage.  Storm surge from hurricane Katrina which appears correct.
possibleOutlierDamage <- which(storm2$totalDamage == max(storm2$totalDamage))
print(t(storm2[possibleOutlierDamage,highlights]))
```

```
##             466257
## EVTYPE      "STORM SURGE"
## BGN_DATE    "8/29/2005 0:00:00"
## STATE       "LA"
## COUNTYNAME  "LAZ040 - 059 - 061>064 - 067>070"
## BGN_LOCATI  ""
## FATALITIES  "0"
## INJURIES    "0"
## totalDamage "3.13e+10"
## REMARKS     "Storm surge damage in southeast Louisiana, especially in the New Orleans area and the c
```

```
# Event with maximum number of people harmed.  Witchita, TX tornado which appears correct.
possibleOutlierHarmed <- which(storm2$totalHarmed == max(storm2$totalHarmed))
print(t(storm2[possibleOutlierHarmed,highlights]))
```

```
##             603050
## EVTYPE      "TORNADO"
## BGN_DATE    "4/10/1979 0:00:00"
## STATE       "TX"
## COUNTYNAME  "WICHITA"
## BGN_LOCATI  ""
## FATALITIES  "42"
## INJURIES    "1700"
## totalDamage "2.5e+08"
## REMARKS     ""
```

Aggregation

Aggregations were calculated (see Data Processing above). In the initial review, certain types of events appeared infrequently and skewed the mean analysis. So for the overall mean calculations a minumum count threshold of 20 was used and for the mean top decile a threshold of 50 was used (so at least 5 are included in the top decile). This ensures that the means are not skewed by a small number of events. These events are still included for the total and max aggregations in case they are still significant even as a small number of events.

```
# Function to rank based on damages and people harmed, append counts by event type.
# Function also accepts a minumum count for an event type to apecific events from skewing the mean calc
# Parameters:
#   agg = the aggregated data frame that will be augmented
#   lab = the label of the type aggregation: Total, Max, Mean Overall or Mean Top Decile
```

```
#   minCount = minumum count of event type which should be included in result
augment <- function(agg, lab, minCount) {
  agg <- merge(agg, counts, by = "EVTYPE" )
  agg <- agg[agg$n >= minCount,]
  agg$rankDamage = factor(rank(desc(agg$totalDamage)), ordered = TRUE)
  agg$rankHarmed = factor(rank(desc(agg$totalHarmed)), ordered = TRUE)
  agg$rankFatalities = factor(rank(desc(agg$totalFatalities)), ordered = TRUE)
  agg$label = lab
  agg
}

# Select the measures to be used in the aggregation and analysis. Prepare a list and count the event ty
measures <- storm2[,c("propDmgAmt", "cropDmgAmt", "FATALITIES", "INJURIES", "totalDamage", "totalHarmed
measures$totalFatalities = measures$FATALITIES
evtypeList <- list(EVTYPE = storm2$EVTYPE)
counts <- storm2 %>% count(EVTYPE)

# Aggregate and augment each type of measurement.
totals <- aggregate(measures, evtypeList, sum)
totals <- augment(totals, "Total", minCount = 0)
maxs <- aggregate(measures, evtypeList, max)
maxs <- augment(maxs, "Max", minCount = 0)
means <- aggregate(measures, evtypeList, mean)
means <- augment(means, "Mean Overall", minCount = 20)

# To assess if extreme instances of a storm type can be particularly dangerous, calculate the mean of t
meanTopDecile <- function(x) {mean(x[{qt<-rank(x)/length(x); qt>=0.9}])}
topDecile <- aggregate(measures, evtypeList, meanTopDecile)
topDecile <- augment(topDecile, "Mean Top Decile", minCount = 50)

# Combine all aggregations into a single frame for analysis and plotting
summary <- rbind(totals, maxs, means, topDecile)
countEvents <- prettyNum(dim(storm2)[1], big.mark =",")
countEventTypes <- prettyNum(length(counts$EVTYPE))
startDate <- min(as.Date(storm$END_DATE, format = "%m/%d/%Y"), na.rm = TRUE)
endDate <- max(as.Date(storm$END_DATE, format = "%m/%d/%Y"), na.rm = TRUE)
```

Plotting Functions

Common plotting functions are used to produce consistent graphs for each measure and aggregation.

```
# Function plots the top n storm types for requested aggregation (e.g. Total, Mean, Max) and measure (D
consequencePlot <- function(topN, agType, measure, filr) {
  rank = paste0("rank", measure)
  tot = paste0("total", measure)
  data <- filter(summary, summary[,rank] <= topN & label == agType)
  data$EVTYPE = factor(data$EVTYPE, levels = data[order(data[,rank], decreasing = TRUE), "EVTYPE"])
  plotf <-ggplot(data, aes_string(x = "EVTYPE", y = tot)) +
        geom_bar(colour = "black", fill = filr, stat="identity", size = .5, width = .8) +
        geom_text(aes(label=paste("n=", n)), size = 3.5, hjust = 1.1) +
        coord_flip() +
        theme(plot.title = element_text(size = rel(1))) +
```

```
        labs(title = paste("Top Storm Types by", agType, measure), y = paste(agType, measure), x="Event
}

# Function creates a grid of 4 related plot covering each aggregation for the requested measure ()
gridPlot <- function(topN, measure) {
  plot1 <- consequencePlot(topN, "Total", measure, "red")
  plot2 <- consequencePlot(topN, "Mean Overall", measure, "lightblue")
  grid.arrange(plot1, plot2, ncol=1)
  plot3 <- consequencePlot(topN, "Mean Top Decile", measure, "green")
  plot4 <- consequencePlot(topN, "Max", measure, "orange")
  grid.arrange(plot3, plot4, ncol=1)
}

# Returns the top ranked event type based on aggregation and measure provided
topValue <- function(agType, measure) {
  rank = paste0("rank", measure)
  tot = paste0("total", measure)
  summary[summary[,rank] == 1 & summary$label == agType,c("EVTYPE", tot)]
}
```

Results

902,296 storm events are included in the data source dating from 'r startDate' to 'r endDate'. Each event is categorized into 985 different types of events (EVTYPE). To review prepared data, the following plots and analysis show the most harmful types of storms based on different criteria.

The 3 sections below analyze each of the following measurement criteria:

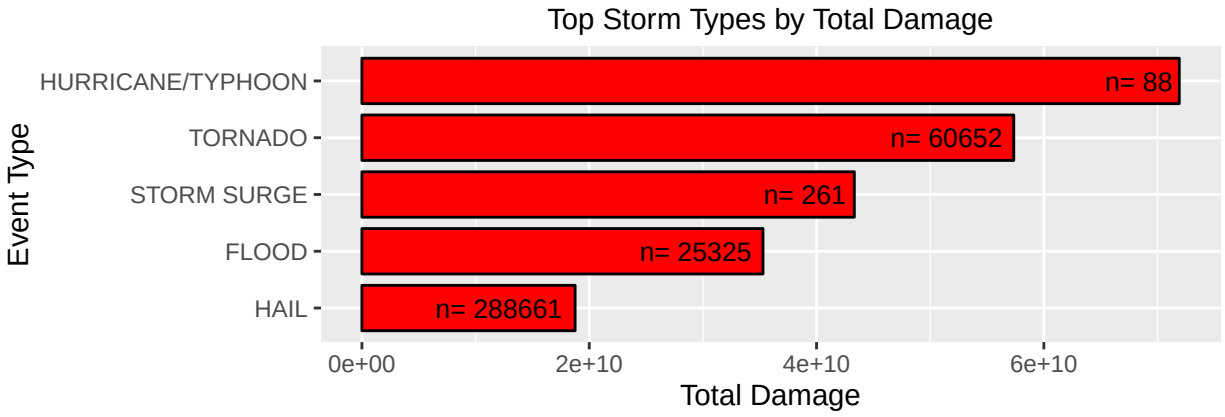
1. Damages ($)
2. People Harmed (Injured + Fatalities)
3. People Killed (Fatalities Only)

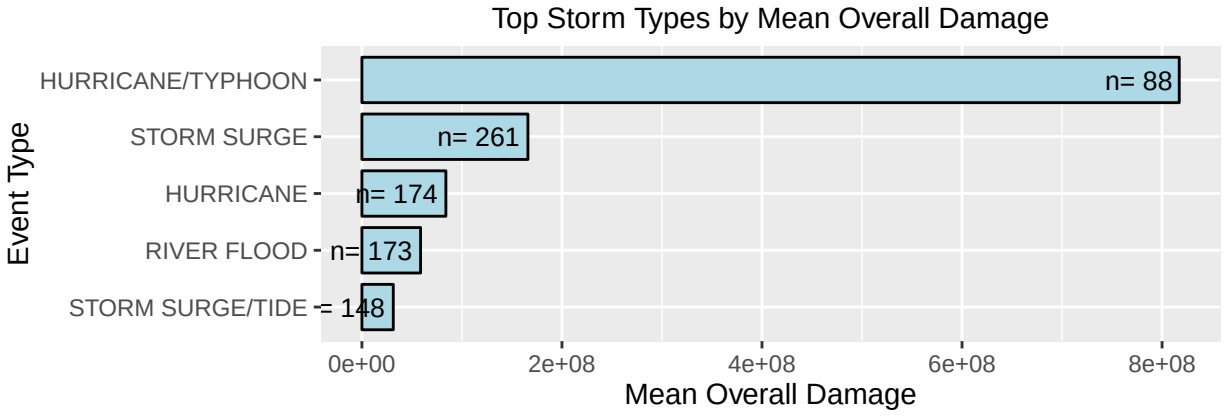
Each section includes a figure with plot grid showing the top 5 ranked results for each type of aggegation. This is followed by summary statistics, highlights of the most significant event and an analysis.

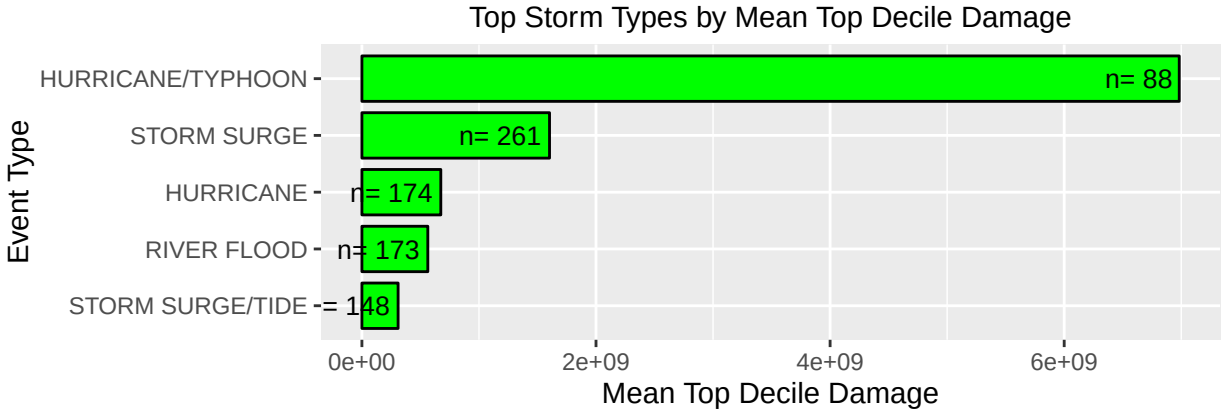
Additional notes on plots:
n = the total number of storms of that type included in the data source. So the Max aggregation includes the value associated with just 1 event (the maximum) and the Mean Top Decile aggregation includes the mean for just the top decile which is approximately n/10 values.

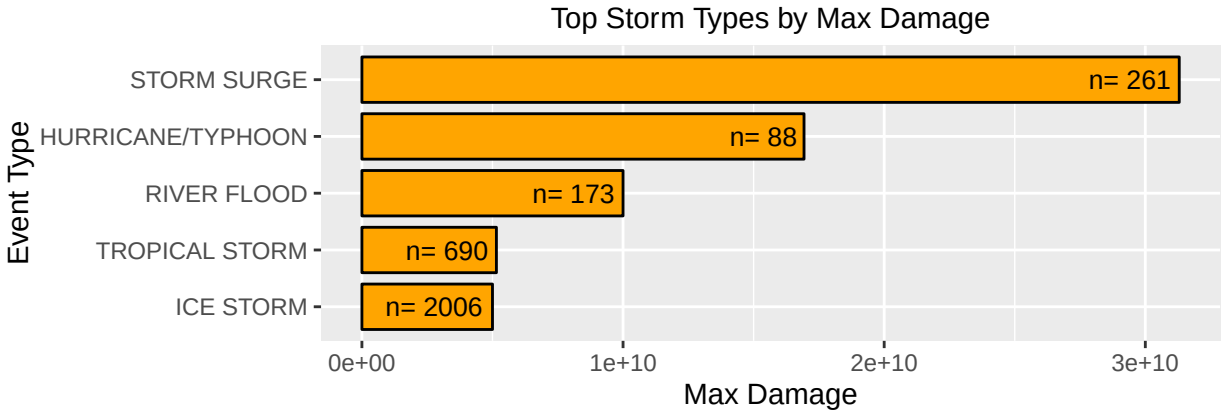
1. Economic Consequences

```
topN = 5
gridPlot(topN, measure = "Damage")
```


Top Storm Types by Total Damage
Event Type
HURRICANE/TYPHOON
TORNADO
STORM SURGE
FLOOD
HAIL
n= 88
n= 60652
n= 261
n= 25325
n= 288661
0e+00
2e+10
4e+10
6e+10
Total Damage


Top Storm Types by Mean Overall Damage
Event Type
HURRICANE/TYPHOON
STORM SURGE
HURRICANE
RIVER FLOOD
STORM SURGE/TIDE
n= 88
n= 261
n= 174
n= 173
= 148
0e+00
2e+08
4e+08
6e+08
8e+08
Mean Overall Damage


Top Storm Types by Mean Top Decile Damage
Event Type
HURRICANE/TYPHOON
STORM SURGE
HURRICANE
RIVER FLOOD
STORM SURGE/TIDE
n= 88
n= 261
n= 174
n= 173
= 148
0e+00
2e+09
4e+09
6e+09
Mean Top Decile Damage


Top Storm Types by Max Damage
Event Type
STORM SURGE
HURRICANE/TYPHOON
RIVER FLOOD
TROPICAL STORM
ICE STORM
n= 261
n= 88
n= 173
n= 690
n= 2006
0e+00
1e+10
2e+10
3e+10
Max Damage

```r
# Top storm type based on total economic impact
topDamage <- topValue("Total", "Damage")
damages <- prettyNum(as.numeric(topDamage[1,"totalDamage"]), big.mark = ",")
print(topDamage)
```

```
##                EVTYPE totalDamage
## 411 HURRICANE/TYPHOON 71913712800
```

```r
# Single most devastating storm in terms of economic impact
topDamageStorm <- storm2[which(storm2$totalDamage == max(measures$totalDamage)),]
print(t(topDamageStorm[,highlights]))
```

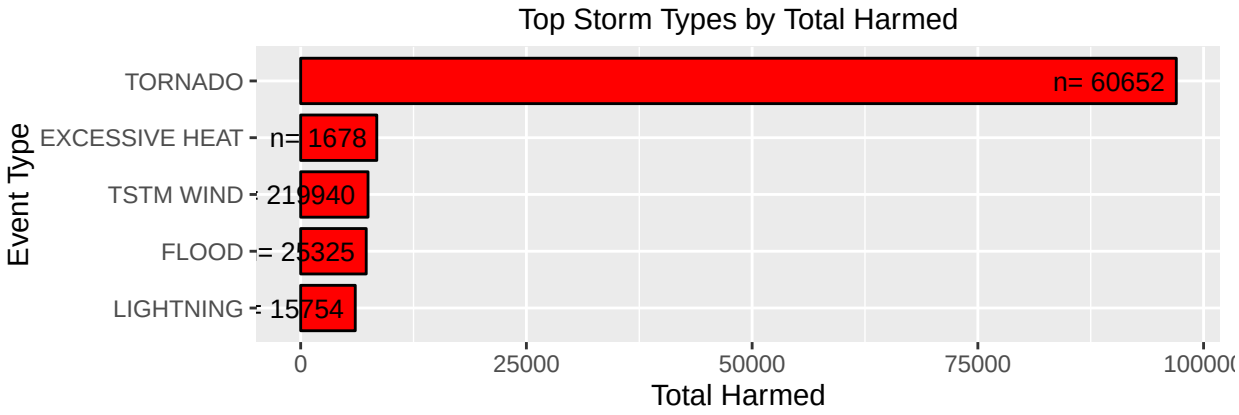
```
##             466257
## EVTYPE      "STORM SURGE"
## BGN_DATE    "8/29/2005 0:00:00"
## STATE       "LA"
## COUNTYNAME  "LAZ040 - 059 - 061>064 - 067>070"
## BGN_LOCATI  ""
## FATALITIES  "0"
## INJURIES    "0"
## totalDamage "3.13e+10"
## REMARKS     "Storm surge damage in southeast Louisiana, especially in the New Orleans area and the c
```

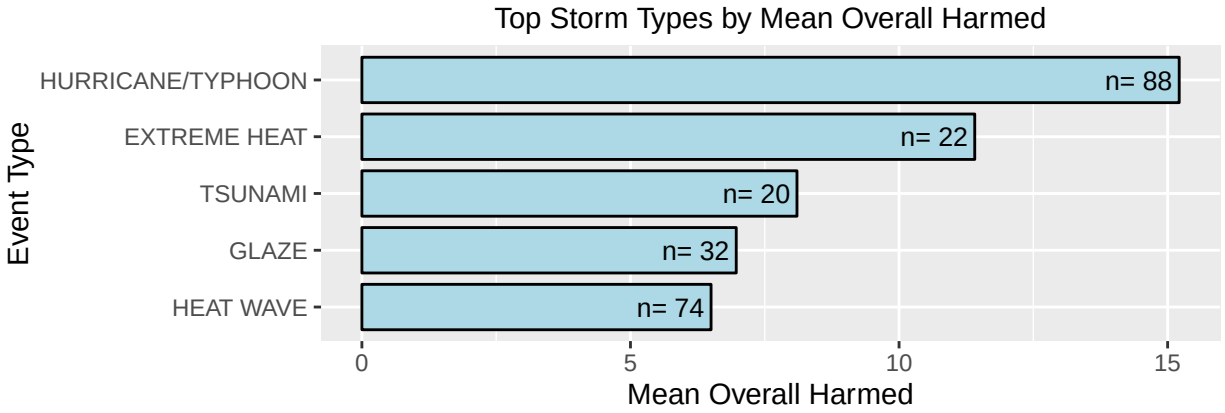
Analysis of Economic Consequences

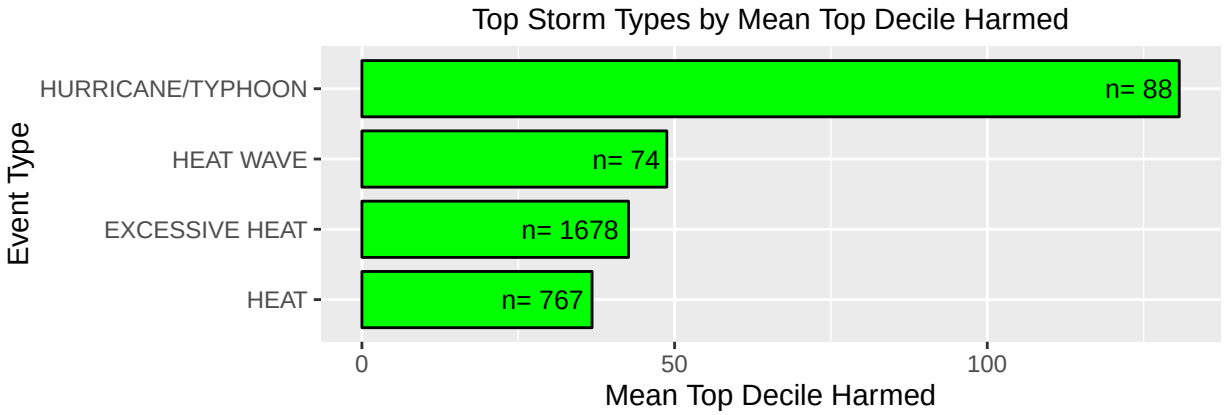
Hurricanes and typhoons are by all aggregate measures the most harmful storm type in terms of economic consequences. The total economic damages caused by hurricanes for all storms included in the data is $71,913,712,800 . The mean impact of the average hurrricane also exceeds any other storm type. Tornados are the next most harmful primarily due the the large number that have occured. Storm surges are the third most harmful storm type, and the storm surge from the 2005 Hurricane Katrina is the most costly on record.

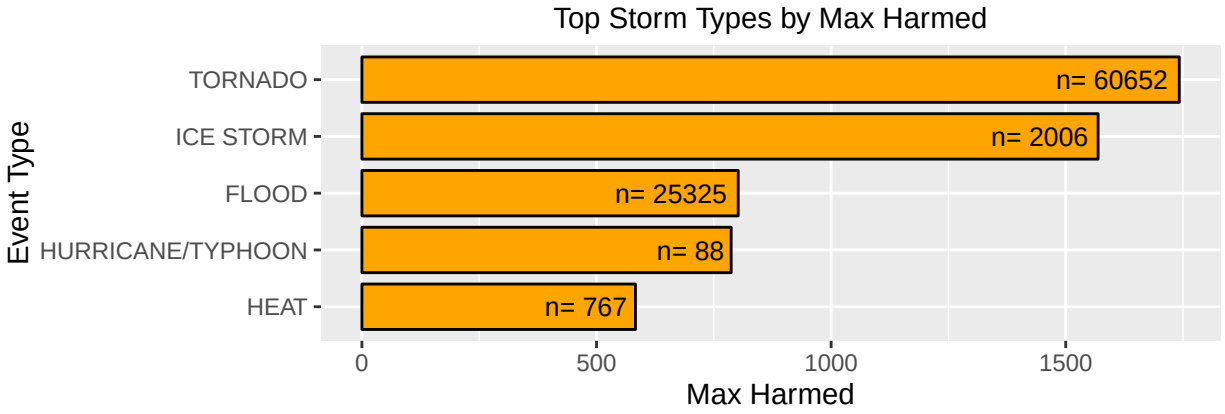
2. People Harmed

```r
gridPlot(topN, measure = "Harmed")
```


Top Storm Types by Total Harmed
TORNADO
EXCESSIVE HEAT
TSTM WIND
FLOOD
LIGHTNING
n= 60652
n= 1678
219940
= 25325
15754
Event Type
0
25000
50000
75000
Total Harmed


Top Storm Types by Mean Overall Harmed
HURRICANE/TYPHOON
EXTREME HEAT
TSUNAMI
GLAZE
HEAT WAVE
n= 88
n= 22
n= 20
n= 32
n= 74
Event Type
0
5
10
15
Mean Overall Harmed


Top Storm Types by Mean Top Decile Harmed
HURRICANE/TYPHOON
HEAT WAVE
EXCESSIVE HEAT
HEAT
n= 88
n= 74
n= 1678
n= 767
Event Type
0
50
100
Mean Top Decile Harmed


Top Storm Types by Max Harmed
TORNADO
ICE STORM
FLOOD
HURRICANE/TYPHOON
HEAT
n= 60652
n= 2006
n= 25325
n= 88
n= 767
Event Type
0
500
1000
1500
Max Harmed

```r
# Top storm type based on number of people hurt or killed
topHarm <- topValue("Total", "Harmed")
harmed <- prettyNum(as.numeric(topHarm[1,"totalHarmed"]), big.mark = ",")
print(topHarm)
```

```
##      EVTYPE totalHarmed
## 834 TORNADO       96979
```

```r
# Top storm type based on mean number of people hurt or killed per storm
topMeanHarm <- topValue("Mean Overall", "Harmed")
meanHarmed <- prettyNum(as.numeric(topMeanHarm[1,"totalHarmed"]), digits = 0)
print(topMeanHarm)
```

```
##                 EVTYPE totalHarmed
## 4111 HURRICANE/TYPHOON    15.21591
```

```r
# Single most devastating storm in terms of number of people hurt or killed
topHarmStorm <- storm2[which(storm2$totalHarmed == max(measures$totalHarmed)),]
print(t(topHarmStorm[highlights]))
```

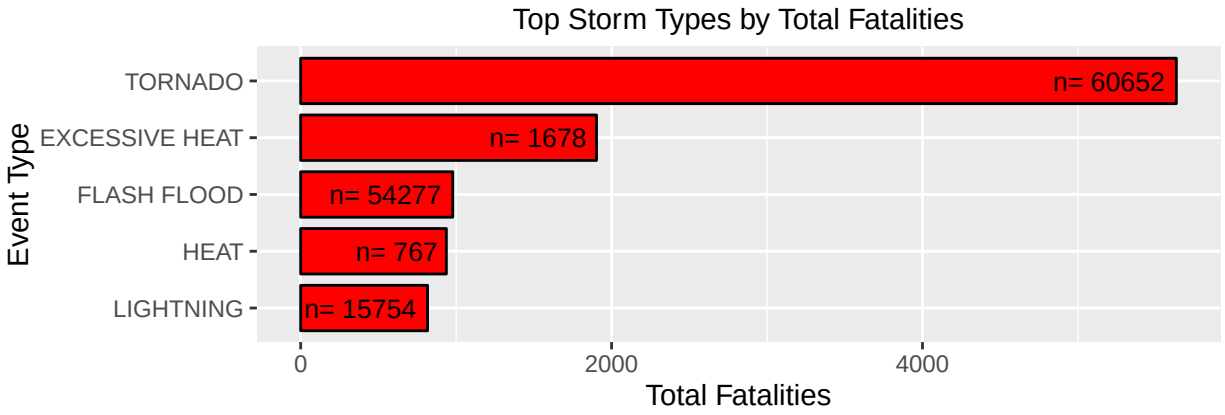
```
##             603050
## EVTYPE      "TORNADO"
## BGN_DATE    "4/10/1979 0:00:00"
## STATE       "TX"
## COUNTYNAME  "WICHITA"
## BGN_LOCATI  ""
## FATALITIES  "42"
## INJURIES    "1700"
## totalDamage "2.5e+08"
## REMARKS     ""
```

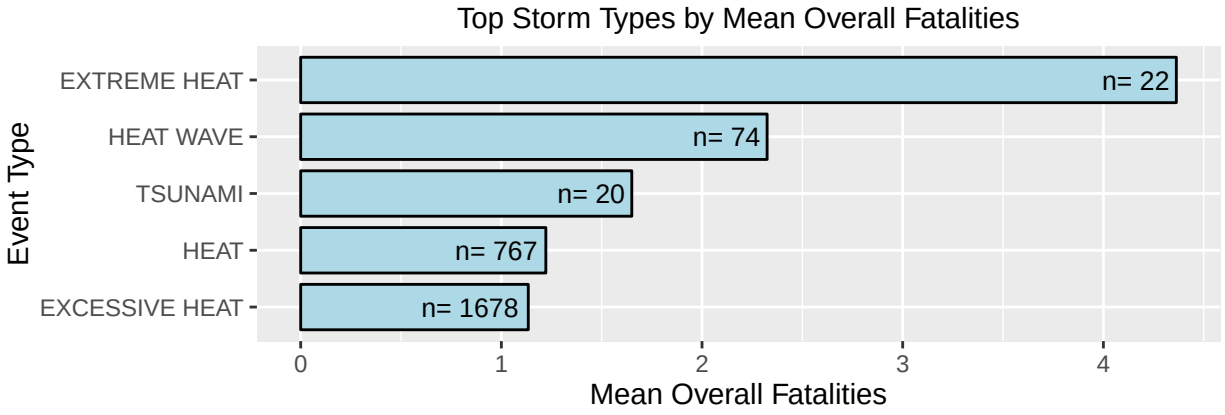
Analysis of People Harmed

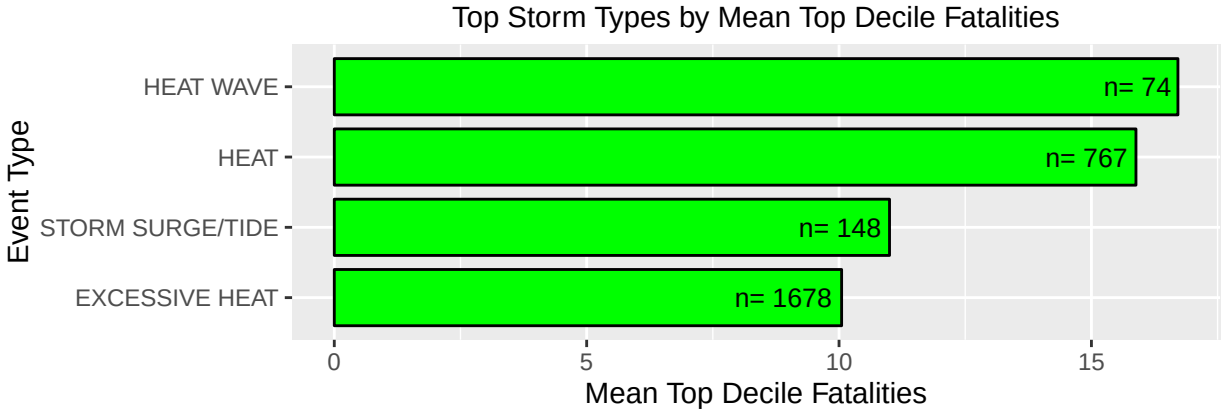
Tornados have hurt the most people of any storm type. 96,979 people have been hurt or killed by tornadoes for all storms tracked in the data source. The average hurricane tends to harm or kill more people (15), but there are many more tornados. The single storm that has caused the most harm to people ever in the US is a Tornado (Wichita, TX in 1979 - see printout).

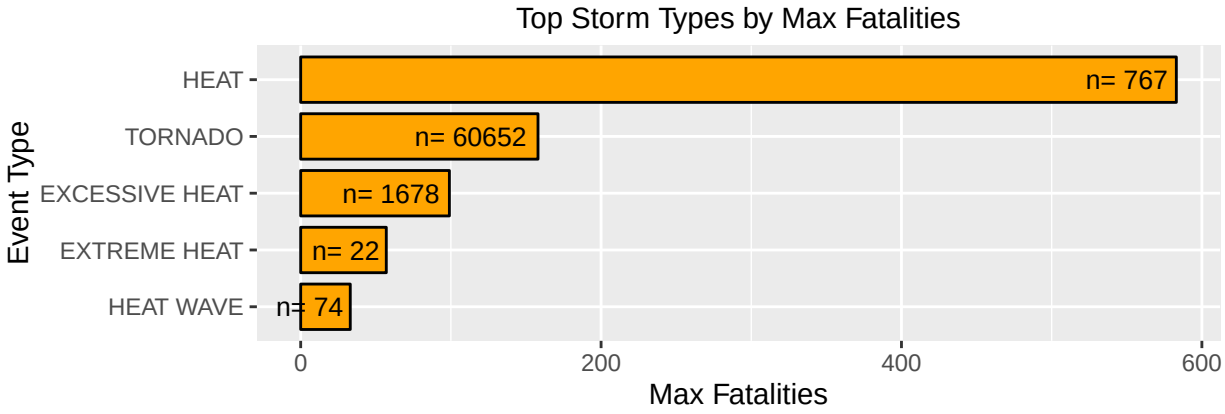
3. Fatalities

```r
gridPlot(topN, measure = "Fatalities")
```


Top Storm Types by Total Fatalities
Event Type
TORNADO
EXCESSIVE HEAT
FLASH FLOOD
HEAT
LIGHTNING
n= 60652
n= 1678
n= 54277
n= 767
n= 15754
0
2000
4000
Total Fatalities


Top Storm Types by Mean Overall Fatalities
Event Type
EXTREME HEAT
HEAT WAVE
TSUNAMI
HEAT
EXCESSIVE HEAT
n= 22
n= 74
n= 20
n= 767
n= 1678
0
1
2
3
4
Mean Overall Fatalities


Top Storm Types by Mean Top Decile Fatalities
Event Type
HEAT WAVE
HEAT
STORM SURGE/TIDE
EXCESSIVE HEAT
n= 74
n= 767
n= 148
n= 1678
0
5
10
15
Mean Top Decile Fatalities


Top Storm Types by Max Fatalities
Event Type
HEAT
TORNADO
EXCESSIVE HEAT
EXTREME HEAT
HEAT WAVE
n= 767
n= 60652
n= 1678
n= 22
n= 74
0
200
400
600
Max Fatalities

```r
# Top storm type based on number of people hurt or killed
topFatalities <- topValue("Total", "Fatalities")
fatalities <- prettyNum(as.numeric(topFatalities[1,"totalFatalities"]), big.mark = ",")
print(topFatalities)
```

```
##      EVTYPE totalFatalities
## 834 TORNADO            5633
```

```r
# Top storm type based on mean number of people hurt or killed per storm
topMeanFatalities <- topValue("Mean Overall", "Fatalities")
meanFatalities <- prettyNum(as.numeric(topMeanFatalities[1,"totalFatalities"]), digits = 1)
print(topMeanFatalities)
```

```
##             EVTYPE totalFatalities
## 14210 EXTREME HEAT        4.363636
```

```r
# Single most devastating storm in terms of number of people killed
deadlyStorm <- storm2[which(storm2$FATALITIES == max(measures$totalFatalities)),]
deadlyStorm$REMARKS = strtrim(deadlyStorm$REMARKS, 1000)
print(t(deadlyStorm[,highlights]))
```

```
##             294149
## EVTYPE      "HEAT"
## BGN_DATE    "7/12/1995 0:00:00"
## STATE       "IL"
## COUNTYNAME  "ILZ003>006 - 008 - 010>014 - 019>023 - 032 - 033 - 039"
## BGN_LOCATI  "Northeast Illinois"
## FATALITIES  "583"
## INJURIES    "0"
## totalDamage "0"
## REMARKS     "An intense heat wave affected northern Illinois from Wednesday July 12 through Sunday Ju
```

Analysis of Fatalities

Tornados are also responsible for the most fatalities. 5,633 people have been killed by tornadoes for all storms tracked in the data source. A extreme heat or a heat wave is most likely to kill someone with a mean number of deaths for extreme heat of 4.

Visualization Data

The summary aggregated data along with some summary statistics is saved to provided data for the shine visualization application.

```r
save(summary, countEvents, countEventTypes, startDate, endDate,
     topDamageStorm, topHarmStorm, deadlyStorm, file = "stormSummary.RData")
```