# The Most Damaging Sever Weather Events: Health and Economic Effects in US (1950-2011)

sha-naya

6/27/2020

## Synopsis

This analysis uses NOAA Storm Database to show which severe weather events have had the most social health and economic impacts on the US population. The data used ranged from 1950 to 2011. The findings indicate that when it comes to population health, that is combined cases of injuries and fatalities caused, tornadoes are the most harmful. In terms of economic consequences, a metric combining crop damage with property damage in USD, floods are the most harmful.

## Data Processing

### Data transformations and Justifications

The raw data was downloaded from this [link] (https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2). Originally, the data contained 902,297 observations of 37 variables. However, the decision was made to edit and transform the data in order to develop a better analysis. Specifically, the manipulations were performed on the PROPDMGEXP and CROPDMGEXP variables. These variables, judging from the documentation [link] (https://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2Fpd01016005curr.pdf) indicate the unit of measurement for PROPDMG and CROPDMG, respectively. For example, "k" stood for thousands and "m" stood for millions. Therefore, if CROPDMG was 5.67 and CROPDMGEXP was "k", it meant that the damage in USD was 5,670 USD. Besides the easily identifieable measurements, like "h", "k", "m", and "b", some unknown classifications were present, such as numbers and symbols like "?". As a result, the decision was made to exclude rows where CROPDMGEXP and PROPDMGEXP variables were none of the "h", "k", "m", and "b". Additionally, for the sake of uniformity in the variables, these characters were all turned to lower case. In total, 971 rows were removed under the specified criteria, or 0.108%. However, more transformations were required. First of all, these characters now needed to become appropiate numbers, so that it would be possible to combine CROPDMG with CROPDMGEXP and PROPDMG with PROPDMGEXP to have two variables that would report the crop and property damage in actual USD values. Thus, CROPDMGEXP and PROPDMGEXP rows with "h" became 100, "k" turned into 1000, and so on. All of the values of CROPDMGEXP and PROPDMGEXP that were NA, were simply turned to 0. Finally, two new variables were created from all of these transformations, crop.dmg.usd and prop.dmg.usd, results of multiplying now transformed pair of variables, CROPDMG with CROPDMGEXP and PROPDMG with PROPDMGEXP. These two new variables now reported the actual damage values in USD.

Please refer to the code below for actual actions performed.

```
storm.data <- read.csv("repdata-data-StormData.csv.bz2")
storm.data$PROPDMGEXP <- tolower(storm.data$PROPDMGEXP)
```

```r
storm.data$CROPDMGEXP <- tolower(storm.data$CROPDMGEXP)
remove.list <- c("-", "?", "+", "0", "1", "2", "3", "4", "5", "6", "7", "8")
for(s in remove.list){
        storm.data <- storm.data[!(storm.data$CROPDMGEXP == s),]
}
for(s in remove.list){
        storm.data <- storm.data[!(storm.data$PROPDMGEXP == s),]
}

storm.data$CROPDMGEXP[storm.data$CROPDMGEXP == "h"] <- "100"
storm.data$CROPDMGEXP[storm.data$CROPDMGEXP == "k"] <- "1000"
storm.data$CROPDMGEXP[storm.data$CROPDMGEXP == "m"] <- "1000000"
storm.data$CROPDMGEXP[storm.data$CROPDMGEXP == "b"] <- "1000000000"
storm.data$CROPDMGEXP <- as.numeric(storm.data$CROPDMGEXP)
storm.data$CROPDMGEXP[is.na(storm.data$CROPDMGEXP)] <- 0

storm.data$PROPDMGEXP[storm.data$PROPDMGEXP == "h"] <- "100"
storm.data$PROPDMGEXP[storm.data$PROPDMGEXP == "k"] <- "1000"
storm.data$PROPDMGEXP[storm.data$PROPDMGEXP == "m"] <- "1000000"
storm.data$PROPDMGEXP[storm.data$PROPDMGEXP == "b"] <- "1000000000"
storm.data$PROPDMGEXP <- as.numeric(storm.data$PROPDMGEXP)
storm.data$PROPDMGEXP[is.na(storm.data$PROPDMGEXP)] <- 0

storm.data$crop.dmg.usd <- storm.data$CROPDMG * storm.data$CROPDMGEXP
storm.data$prop.dmg.usd <- storm.data$PROPDMG * storm.data$PROPDMGEXP

event.list <- storm.data$EVTYPE
event.list <- event.list[!duplicated(event.list)]
```

## Results

The analysis aimed to answer two questions. First, across the United States, which types of events are most harmful with respect to population health? Second, across the United States, which types of events have the greatest economic consequences? Population health, in this scenario, was defined as the combined number of fatalities and injuries sustained during any given event. It was decided to take the sum of fatalities and injuries instead of the mean, over the 1950-2011 period, because the question asked which events were the most harmful, not most harmful on average.

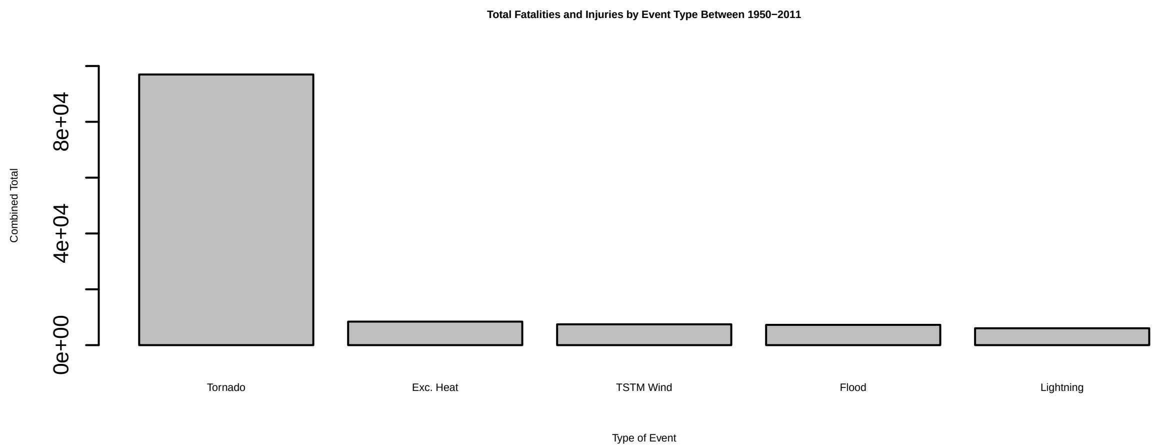
Please see the code below to see how final results were achieved in technical terms.

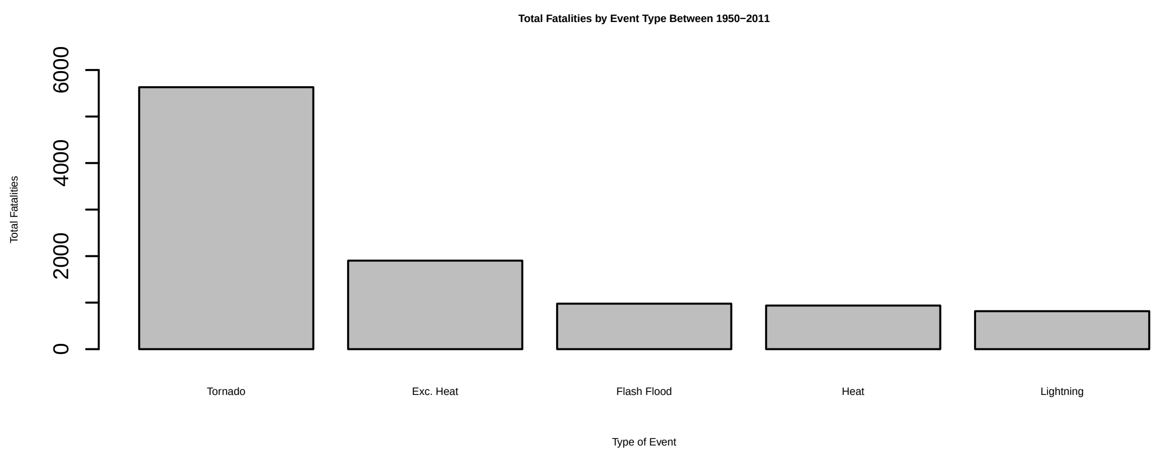
```r
sum.event.ph <- data.frame(EventType = character(),
                          FatalitiesTotal = numeric(), InjuriesTotal = numeric())
for(e in event.list) {
      df1 <- subset(storm.data, EVTYPE == e)
      sum1 <- sum(df1$FATALITIES)
      sum2 <- sum(df1$INJURIES)
      df2 <- data.frame(EventType = e,
                        FatalitiesTotal = sum1, InjuriesTotal = sum2)
      sum.event.ph <- rbind(sum.event.ph, df2)
}

sum.event.ph.f <- sum.event.ph[order(-sum.event.ph$FatalitiesTotal),]
sum.event.ph.f <- sum.event.ph.f[1:5,]
```

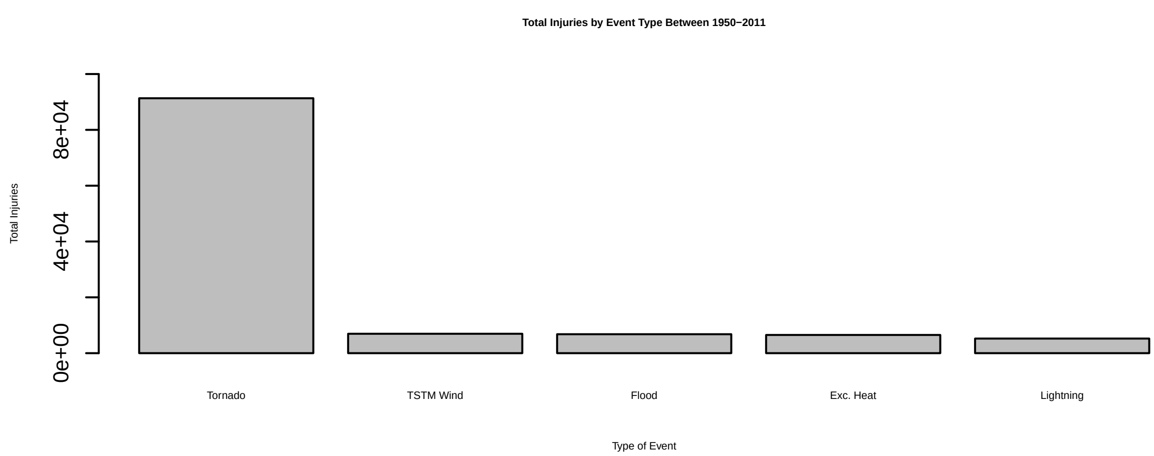
```r
sum.event.ph.i <- sum.event.ph[order(-sum.event.ph$InjuriesTotal),]
sum.event.ph.i <- sum.event.ph.i[1:5,]

sum.event.ph$CombinedTotal <- sum.event.ph$FatalitiesTotal + sum.event.ph$InjuriesTotal
sum.event.ph.c <- sum.event.ph[order(-sum.event.ph$CombinedTotal),]
sum.event.ph.c <- sum.event.ph.c[1:5,]
```

The event with the most fatalities is the "Tornado". The one with the most injuries is again the "Tornado". As a result, the event with the most damaging effect of health population is the "Tornado", which accounted for a combined of 96,951 cases (fatalities and injuries). The top five for each category are provided in the below barplot figure.

```r
par(mfrow=c(3,1))
barplot(sum.event.ph.c$CombinedTotal, ylim = c(0, 100000), xlab = "Type of Event",
        names.arg = c("Tornado", "Exc. Heat", "TSTM Wind", "Flood", "Lightning"),
        cex.names = 0.5, ylab = "Combined Total", cex.lab = 0.5,
        main = "Total Fatalities and Injuries by Event Type Between 1950-2011", cex.main = 0.5)
barplot(sum.event.ph.f$FatalitiesTotal, ylim = c(0, 6000), xlab = "Type of Event",
        names.arg = c("Tornado", "Exc. Heat", "Flash Flood", "Heat", "Lightning"),
        cex.names = 0.5, ylab = "Total Fatalities", cex.lab = 0.5,
        main = "Total Fatalities by Event Type Between 1950-2011", cex.main = 0.5)
barplot(sum.event.ph.i$InjuriesTotal, ylim = c(0, 100000), xlab = "Type of Event",
        names.arg = c("Tornado", "TSTM Wind", "Flood", "Exc. Heat", "Lightning"),
        cex.names = 0.5, ylab = "Total Injuries", cex.lab = 0.5,
        main = "Total Injuries by Event Type Between 1950-2011", cex.main = 0.5)
```




Economic consequence was defined as the sum of crop and property damage in USD, that is, a sum of crop.dmg.usd and prop.dmg.usd variables. Again, it was decided to take the sum of these values instead of the average given the wording of the question.

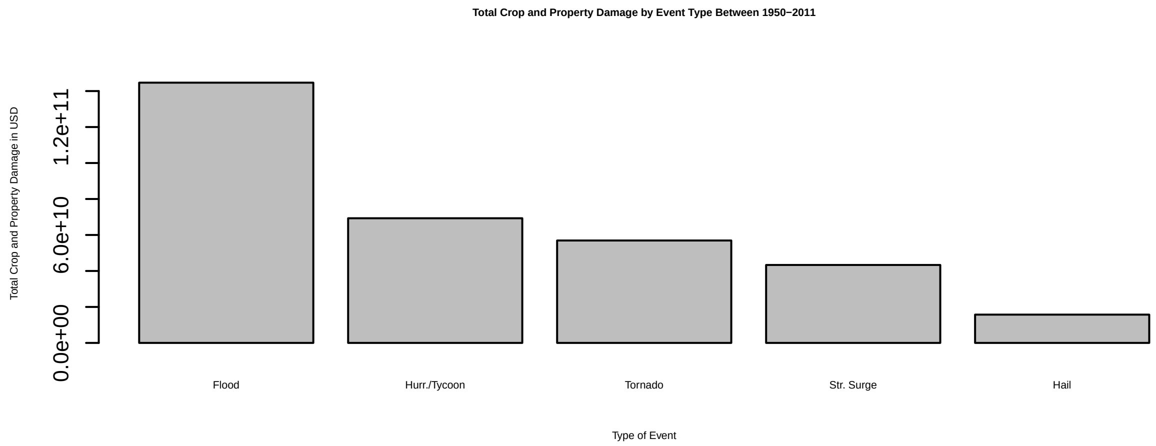
The code below shows the technical actions performed to achieve the results.

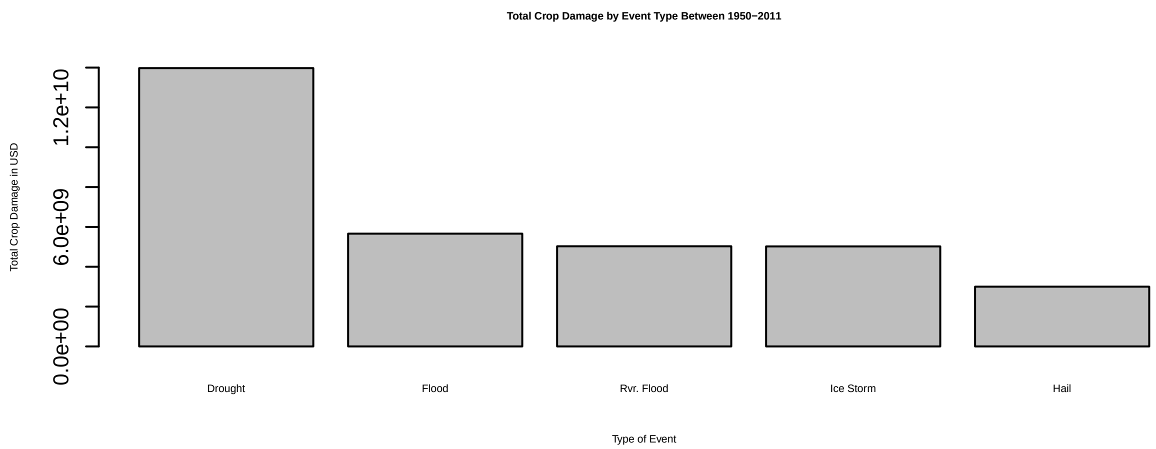
```
sum.event.ec <- data.frame(EventType = character(),
                           CropDMGTotal = numeric(), PropDMGTotal = numeric())
```

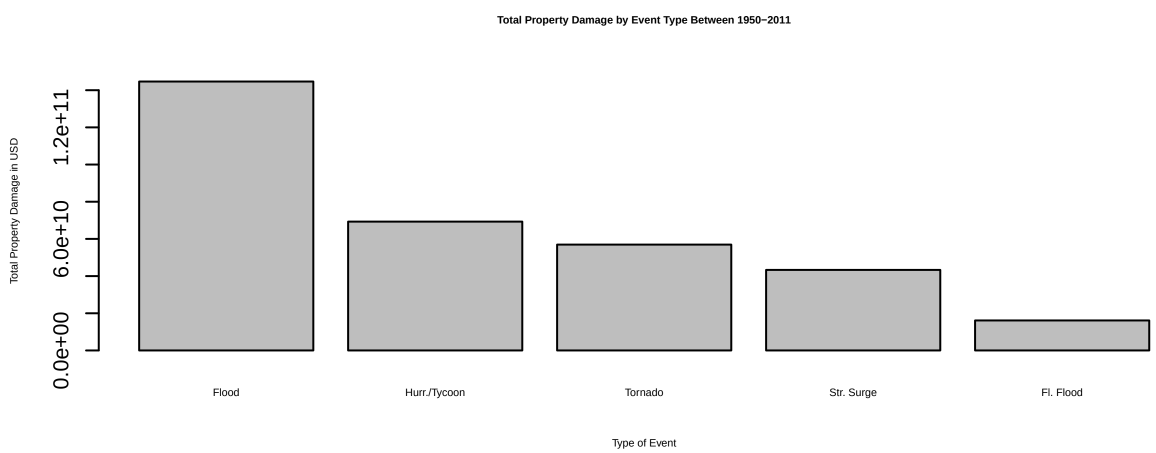
```r
for(e in event.list) {
        df3 <- subset(storm.data, EVTYPE == e)
        sum3 <- sum(df3$crop.dmg.usd)
        sum4 <- sum(df3$prop.dmg.usd)
        df4 <- data.frame(EventType = e,
                           CropDMGTotal = sum3, PropDMGTotal = sum4)
        sum.event.ec <- rbind(sum.event.ec, df4)
}

sum.event.ec.c <- sum.event.ec[order(-sum.event.ec$CropDMGTotal),]
sum.event.ec.c <- sum.event.ec.c[1:5,]

sum.event.ec.p <- sum.event.ec[order(-sum.event.ec$PropDMGTotal),]
sum.event.ec.p <- sum.event.ec.p[1:5,]

sum.event.ec$CombinedTotal <- sum.event.ec$CropDMGTotal + sum.event.ec$PropDMGTotal
sum.event.ec.b <- sum.event.ec[order(-sum.event.ec$CombinedTotal),]
sum.event.ec.b <- sum.event.ec.b[1:5,]
```

The event with the most damaging effect measured in only crop damages is "Drought". When it comes to only measuring property damage, the event is "Flood". Finally, the event with most damaging economic consequences, crop and property damage combined, is "Flood" with 150,319,678,250USD in damages. The top five are provided below in the barplot figure.

```r
par(mfrow=c(3,1))
barplot(sum.event.ec.b$PropDMGTotal, ylim = c(0, 155000000000), xlab = "Type of Event",
        names.arg = c("Flood", "Hurr./Tycoon", "Tornado", "Str. Surge", "Hail"),
        cex.names = 0.5, ylab = "Total Crop and Property Damage in USD", cex.lab = 0.5,
        main = "Total Crop and Property Damage by Event Type Between 1950-2011",
        cex.main = 0.5)
barplot(sum.event.ec.c$CropDMGTotal, ylim = c(0, 14000000000), xlab = "Type of Event",
        names.arg = c("Drought", "Flood", "Rvr. Flood", "Ice Storm", "Hail"),
        cex.names = 0.5, ylab = "Total Crop Damage in USD", cex.lab = 0.5,
        main = "Total Crop Damage by Event Type Between 1950-2011", cex.main = 0.5)
barplot(sum.event.ec.p$PropDMGTotal, ylim = c(0, 150000000000), xlab = "Type of Event",
        names.arg = c("Flood", "Hurr./Tycoon", "Tornado", "Str. Surge",
        "Fl. Flood"), cex.names = 0.5, ylab = "Total Property Damage in USD",
        cex.lab = 0.5, main = "Total Property Damage by Event Type Between 1950-2011",
        cex.main = 0.5)
```




# Conclusion

This analysis aimed to answer two questions in regard to which type of severe weather causes the most effect on population health and which one is responsible for most damaging economic consequences. Ultimately, however, this analysis is very primitive and whoever finds this useful must first take into consideration any new severe weather developments in the past 9 years (the data used for this analysis ranged only from 1950 to 2011) and also consider the importance of once's geographical location in the US (some regions in the country are more prone to specific severe weather occurences). A more sophisticated analysis would answer the two questions by state or region criteria, rather than about the whole of the United States. Another useful modification could include a time series plot describing the change over time in some of these severe weather events.

## Notes

The original .R file could be provided upon request (please see my GitHub: sha-naya) Below you will find my system specifications and other relevant information.

```r
sessionInfo()
```

```
## R version 4.0.0 (2020-04-24)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Catalina 10.15.5
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.0.0  magrittr_1.5    tools_4.0.0     htmltools_0.4.0
##  [5] yaml_2.2.1      Rcpp_1.0.4.6    stringi_1.4.6   rmarkdown_2.2
##  [9] knitr_1.28      stringr_1.4.0   xfun_0.14       digest_0.6.25
## [13] rlang_0.4.6     evaluate_0.14
```