# Analyzing Impact of US Storms

*Rohit Padebettu*

*10/17/2016*

## Synopsis

*In this project, we analyze the storm data acquired from National Weather Service to assess its impact on human life as well as the economic damages caused by such storms over the last 20 years. The data provided by the National Weather Service goes back to the year 1950, but since all events and event types weren't recorded, cumulating such data would lead to faulty analysis. So here we use the data from the year 1993 onwards to perform our analysis when most of the events started getting recorded*

*We begin by gathering and cleaning up the data to allow analysis. We then proceed to compute the economic damage and human damage caused by such events over the years. We summarize our findings via the top 10 event types that caused the maximum human damage and econoic damage. We also show the economic damages caused by such events for every year since 1993*

*We conclude that Flood and Flood related damages have caused by far the greatest economic damage, especially in the years 2005 and 2006. Tornadoes and Flood related events have also lead to the greatest human damage assesed via total fatalities and deaths caused by each event type. Certain events like flash floods and heat have tendencies to cause a greater number of deaths in relation to the human damages they cause*

## Analysis of Data

We begin our analysis by first loading a set of R libraries which aid us in various steps of our analysis.

```
suppressPackageStartupMessages(library(R.utils))          # for unzipping the file
suppressPackageStartupMessages(library(data.table))       # for quick reading and subsetting
suppressPackageStartupMessages(library(lubridate))        # for date transformations
suppressPackageStartupMessages(library(dplyr) )           # for data transformations
suppressPackageStartupMessages(library(ggplot2))          # for plots
suppressPackageStartupMessages(library(ggthemes))         # themes for plots
suppressPackageStartupMessages(library(plotly) )          # for fancier interactive plots
```

## Data Processing

### Raw Data

We begin our data processing by first obtaining the raw data from the course website which in turn was obtained from the NOAA Website.

The file provided for the course project was 47Mb in size and was in compressed format. We have included below the code we have used to download and uncompress the file via `R.Utils` package.

```
## Downloading file and decompressing
setwd("./Final Course Project/")
url<-c("https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2")

## Run the following code once
download.file(url,"./Data/StormData.csv.bz2",method = "curl")
bunzip2("./Data/StormData.csv.bz2",remove = FALSE)
```

### Reading the data

We read the data into the R session via the `fread` function available in the `data.table` package. We choose this because of the speed and efficiency of the function in reading such a large file into memory.

```
StormData<-fread("./Data/StormData.csv",header = TRUE,strip.white = TRUE)
```

```
##
Read 22.7% of 967216 rows
Read 49.6% of 967216 rows
Read 70.3% of 967216 rows
Read 82.7% of 967216 rows
Read 902297 rows and 37 (of 37) columns from 0.523 GB file in 00:00:07
```

### Date Processing

Once the raw data is read into the memory, we see from the output of the `fread` function that it has *902297* rows and *37* variables. The `data.table` itself takes about 0.523GB space in our memory.We then proceed to transform the date columns in the dataset which appear as `Character` format. We use the `lubridate` package here to transform the `BGN_DATE` and 'END_DATE' columns into Date format. We do this transformation to allow us to subset the data starting from year 1993 onwards.

Though the overall dataset contains events beginning in 1950, recording of such events for all varieties of storms began around 1990's. If we chose to use the whole dataset, our analysis could potentially be skewed by the **bias in data collection**

Furthermore we choose only few of 37 columns available to us in order to reduce the memory footprint and also use only the data required to perform the analysis.

```
StormData$BgnDate <- as_date(mdy_hms(StormData$BGN_DATE))
StormData$EndDate <- as_date(mdy_hms(StormData$END_DATE))

StormData_sub <- StormData[year(BgnDate)>1992,
                           .(BgnDate,EndDate,REFNUM,
                            STATE,STATE__,COUNTY,COUNTYNAME,
                            EVTYPE,FATALITIES,INJURIES,
                            PROPDMG,PROPDMGEXP,CROPDMG,CROPDMGEXP)]

Data_df<-as.data.frame(StormData_sub)
```

### Economic Damage computation

The dataset available to us has two columns `PROPDMG` and `PROPDMGEXP` which together help us get an assessment of total property damage caused by one particular storm/event. Similarly we also have `CROPDMG` and `CROPDMGEXP` which help us understand the Crop damages caused by the storm. To assess the total economic damage, we must compute the total damage caused by the storm to Property as well as Crop.

The variables `PROPDMGEXP`,`CROPDMGEXP` in the dataset are coded factors where `M` stands for a `million`,`B` for a `billion`, `K` for a `Thousand` and so on. So in order to allow us to calculate the total damage, we must first decode these variables into their numeric forms. the function below `decoder` helps us do the same

```
decoder<-function(v){
    if(length(v[2])==0){return (1)}
    else if(v[2]%in% c("H","h")) {return(100)}
    else if(v[2]%in% c("K","k")) {return(1000)}
    else if(v[2]%in% c("M","m")) {return(1000000)}
    else if(v[2]%in% c("B","b")) {return(1000000000)}
    else if(v[2]%in% c("0","1","2","3","4","5","6","7","8","9")) {10^as.numeric(v[2])}
    else if(v[2]=="") {return(1)}
    else return(1)
}
```

### Preparing a clean dataset

In the code give below, we apply the function defined above to the variables and create a new column `Tot_damage` which gives us the Total economic damage caused by the event to both property and crop.

This dataset we prepare here will be used for the various downstream analysis we will perform

```
PROPEXP<-apply(Data_df[,c(11,12)],1,decoder)
CROPEXP<-apply(Data_df[,c(13,14)],1,decoder)
Data_df<-cbind(Data_df,PROPEXP,CROPEXP)
Data_df <- mutate(Data_df,Tot_damage = (PROPDMG*PROPEXP)+(CROPDMG*CROPEXP))
```

## Results

### Event Types which cause the most economic damage

The first analysis we perform is to find out which of the event types in the dataset have caused the highest economic damage.

Since most of the events in the dataset seem to have caused relatively a small amount of economic damage, we prune down the events set to include only the events which are in the top 50% of damage caused.We proceed to clean up the EVTYPE variable to a reasonable extent to allow us to factorize the variable and aggregate the data over each event type.

*It still appears from the resulting* ***Event*** *factors that there are potentially duplicate Event Types still encoded separately, but cleaning those up will require a lot more manual effort and time which is beyond the scope of the current analysis.*

```
n=50
Data_calc<-Data_df%>%
    filter(Tot_damage > quantile(Tot_damage, prob = 1 - n/100))%>%
    select(Date=BgnDate,Event=EVTYPE ,Damage = Tot_damage)%>%
    mutate(Event = as.factor(toupper(trim(Event))))%>%
    group_by(Event)%>%
    summarize(EventDamage = sum(Damage)/1000000000)%>%
    arrange(desc(EventDamage))
```

### Top 10 Event Types causing the most economic damage

Once the data is aggregated and sorted as above, we proceed to select the top 10 event types from the resulting list and plot them against one another as below.

As we can see below the maximum economic damage has been caused by Flood related events by far, followed by Hurricane, Storm and Tornado events.

```
Data_top10<-Data_calc[c(1:10),]

g1<-ggplot(Data_top10,aes(x=factor(Event),y=EventDamage,fill=Event))+
    geom_bar(stat="identity",col="black",width=0.5)+
    coord_flip()+
    guides(fill=FALSE)+
    ylab("Economic Damage(in $ billions)")+
    xlab("Top 10 Event Types")+
    ggtitle("Most Economically Damaging Event Types")+
    theme_gdocs()
print(g1)
```

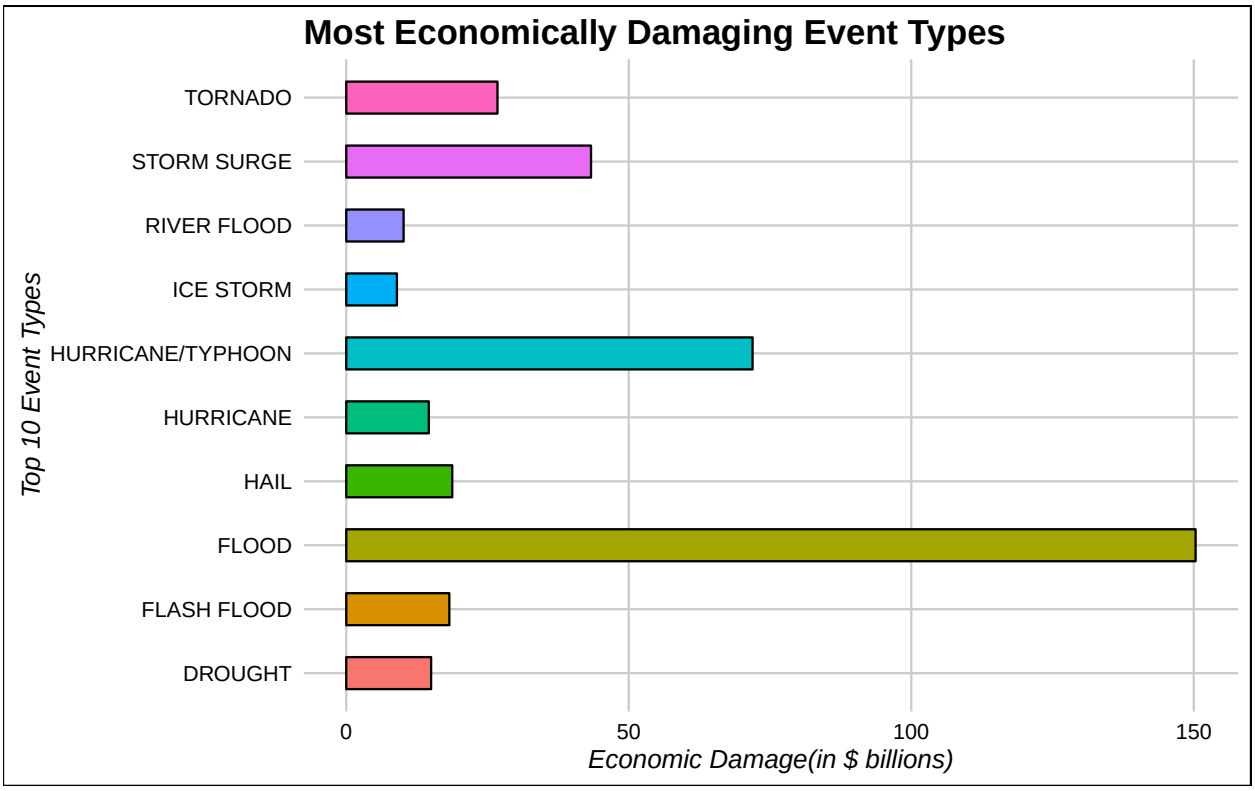


Figure 1: Top 10 most economically damaging events

### Yearly Economic Damage

In the second part of our analysis, we try to understand the economic damage caused by these weather events on a yearly basis.

Our cleaned dataset allows us to do this analysis quickly, where we just have to aggregate the data by `Year` (derived from the Date column) instead of `Event`. We do the aggregation and sorting in the code below using the `dplyr` package as in the last section

```
Data_year<-Data_df%>%
    select(Date=BgnDate,Event=EVTYPE,Damage = Tot_damage)%>%
    group_by(Year = year(Date))%>%
    summarize(YearlyDamage = sum(Damage)/1000000000)%>%
    arrange(Year)
```

When we plot the year over year economic damage caused by the storms we can clearly see from the chart below that the maximum economic damage was caused in the years **2005** and **2006** mostly related to flooding and displacement caused by the **Katrina hurricane**

```
g2<-ggplot(Data_year,aes(x=Year,y=YearlyDamage))+
    geom_bar(stat="identity",col="black",width=0.5,fill='salmon')+
    ylab("Economic Damage(in $ billions)")+
    xlab("Years")+
    ggtitle("Yearly Economic Damage")+
    theme_gdocs()
print(g2)
```

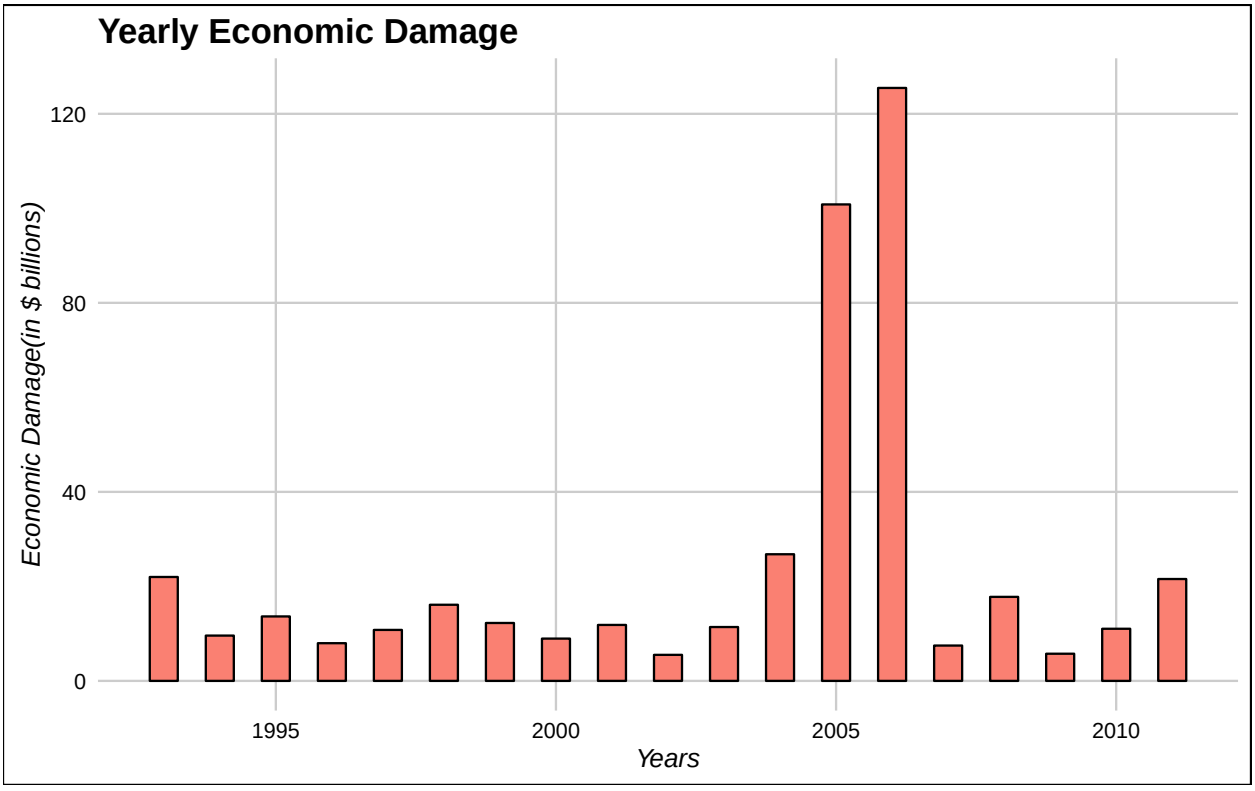


Figure 2: Yearly Economic Damage from weather events

### Event Types which cause the most Human Damage

In the last part of our analysis we try and understand the human impact of these weather events. We analyze both the Fatalities and Injuries reported in the dataset to come up with two new variables: `HumanDamage` which we define as the total of injuries and fatalities caused by each event and `Intensity` which we define as the ratio of fatalities to `HumanDamage`.

This allows us to categorize the event types not just by how many people were impacted by each event, but also understand how hard the impact was per event type.We use the cleaned up dataset to select the `FATALITIES` and `INJURY` variables. As done in the first economic damage analysis, we clean and factorize the `EVTYPE` data into `Event` to allow us to aggregate and sort

```
Data_hum<-Data_df%>%
    select(Date=BgnDate,Event=EVTYPE,FATALITIES,INJURIES)%>%
    mutate(Event = as.factor(toupper(trim(Event))))%>%
    group_by(Event)%>%
    summarize(HumanDamage=sum(FATALITIES,INJURIES),
              Deaths=sum(FATALITIES))%>%
    mutate(Intensity = Deaths/HumanDamage)%>%
    arrange(desc(HumanDamage))
```

### Top 10 Event Types causing most Human Damage

After we aggregate the data and calculate our new `HumanDamage` and `Intensity` variables per event type, we sort them in decreasing order of `HumanDamage`. We then pick the top 10 events from the resulting set to plot below. The chart below is also color coded to show the `Intensity` of each event type (with darker colors showing higher intensity).

We can clearly see below that while Tornadoes and Flood related events have caused by far the greated human damage. Flash Floods and Heat related events cause 3 times as many deaths among the reported cases.

```
Data_top10_h<-Data_hum[c(1:10),]

g3<-ggplot(Data_top10_h,aes(x=factor(Event),y=HumanDamage,fill=Intensity))+
    geom_bar(stat="identity",col="black",width=0.7)+
    scale_fill_gradient(low = "yellow", high = "red")+
    coord_flip()+
    ylab("Human Damage(Fatalities+Injuries)")+
    xlab("Top 10 Event Types")+
    ggtitle("Event types causing most human damage")+
    theme_fivethirtyeight()
print(g3)
```

## Conclusion

From our analysis of the US Storm dataset for the period starting in 1993,we conclude the following:

- Flood and Flood related damages have caused by far the greatest economic damage.
- The greatest economic damage in the last 20 years occurred in the years 2005 and 2006 as a result of Hurrican Katrina.
- Tornadoes and Flood related events have also lead to the greatest human damage.
- Heat and Flash Flood related events seem to have some of the highest intensities of death in relation to the human damage caused.

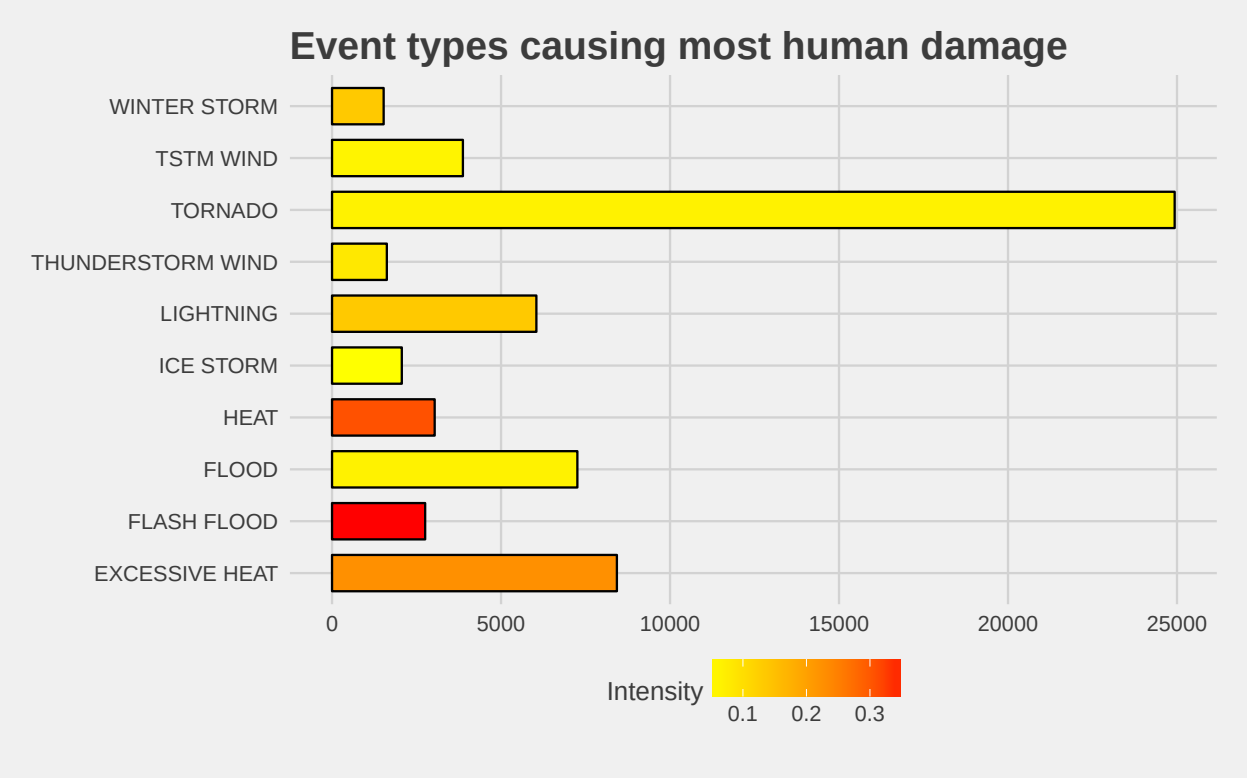


Figure 3: Top 10 weather events most harmful to humans