Analysis of historical storm events and its impact on economic and population health

Efrain Villalobos

5/17/2020

github repository (https://github.com/scrain/reproducible-research-project-2)

Synopsis

U.S. National Oceanic and Atmospheric Administration’s (NOAA) storm database begain tracking a standard set of 48 storm data events in 1996. After analyzing storm data events from 1996 to 2011 it was found that Hurricanes/Typhoons cause the most economic impact in relation to crop and property damage, while Tornados take the most population toll in regards to injuries and fatalities.

Table of Contents

Data Processing

Retrieval and Loading

The compressed data is conditionally downloaded from the source URL if not found locally and then loaded directly via `read.csv` . Before proceeding, some basic validation is done on the file and dataset per some advice found from a course mentor in the discussion forums here (https://www.coursera.org/learn/reproducible-research/discussions/weeks/4/threads/IdtP_JHzEeaePQ71AQUtYw).

```
filename <- 'StormData.csv.bz2'
if (!file.exists(filename)) {
  download.file('https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2', filename)
}
storm_data <- read.csv(filename)
# Ensure we got the data downloaded, decompressed and loaded correctly
# by checking filesize and dataset dimemsions
stopifnot(file.size(filename) == 49177144)
stopifnot(dim(storm_data) == c(902297,37))
```

Cleaning and Preparation

After the dataset is loaded, it is cleaned as follows:

1. Given the poor quality of the property and crop damage exponent variables (`PROPDMGEXP` and `CROPDMGEXP`), two variables are added to hold converted multiplier values named `PropDamageMult` and `CropDamageMult` . Again from advice found in the previously mentioned forum post, an approach is used based on the analysis found in the article “How To Handle Exponent Value of PROPDMGEXP and CROPDMGEXP” (https://rstudio-pubs-static.s3.amazonaws.com/58957_37b6723ee52b455990e149edde45e5b6.html). Using this information, the function `convertExponentToMultiplier` is used to convert the original exponent variables into the corresponding multipliers. See appendix A for the result of this exponent to multiplier conversion.

```
convertExponentToMultiplier <- function(exp) {
  ifelse(
    exp == '+', 1,                         # '+' -> 1
    ifelse(
      exp %in% paste(seq(0,8)), 10^1,      # 0-8 -> 10
      ifelse(
        exp %in% c('H', 'h'), 10^2,        # H,h -> 100
        ifelse(
          exp %in% c('K', 'k'), 10^3,      # K,k -> 1,000
          ifelse(
            exp %in% c('M', 'm'), 10^6,    # M,m -> 1,000,000
            ifelse(
              exp %in% c('B', 'b'), 10^9,  # B,b -> 1,000,000,000
              0                            # everything else -> 0
            )
          )
        )
      )
    )
  )
}
storm_data$PropDamageMult <- convertExponentToMultiplier(storm_data$PROPDMGEXP)
storm_data$CropDamageMult <- convertExponentToMultiplier(storm_data$CROPDMGEXP)
```

2. With the mulitplier variables created, `CropDamage` and `PropDamage` variables are added by multiplying them against the corresponding damage variables `PROPDMG` and `CROPDMG` . In addition, a `TotalDamage` variable is also added, using the sum of both the crop and property damage.

```
storm_data$PropDamage  <- storm_data$PROPDMG * storm_data$PropDamageMult
storm_data$CropDamage  <- storm_data$CROPDMG * storm_data$CropDamageMult
storm_data$TotalDamage <- storm_data$PropDamage + storm_data$CropDamage
```

3. For determining the oveall health impact of events, a `PopulationHealthImpact` variable is added using the sum of `FATALITIES` and `INJURIES` variables.

```
storm_data$PopulationHealthImpact <- storm_data$FATALITIES + storm_data$INJURIES
```

4. To make the dataset easier to work with, irrelevant observations are removed. According to the documentation, it was not until 1996 that all event types were being recorded. For that reason, the years from the dataset earlier than that are removed in order to get a fair assessment of all events. Also, since we are answering questions around economic and population health impact, all rows having neither of these are removed as well.

```
storm_data$BeginDate   <- as.Date(storm_data$BGN_DATE, '%m/%d/%Y')
sd <- storm_data[storm_data$BeginDate >= '1996-01-01',]
sd <- sd[sd$TotalDamage > 0 | sd$PopulationHealthImpact  > 0,]
```

5. Looking at the top events with the most `TotalDamage` and `PopulationHealthImpact` showed that there was a least one event that had far more economic impact than any other. Using the NOAA Storm Events Database (https://www.ncdc.noaa.gov/stormevents/choosedates.jsp?statefips=-999%2CALL), it was found that a 2006 flood in Napa County, Califorina (https://www.ncdc.noaa.gov/stormevents/listevents.jsp?eventType=%28Z%29+Flood&beginDate_mm=01&beginDate_dd=01&beginDate_yyyy=2006&endDate_mm=01&endDate_dd=01&endDate_yyyy=2006&cou was mis-entered with a `PROPDMGEXP` of **B**illion instead of **M**illion. The erroneous `PROPDMGEXP` value was then corrected and the `PropDamageMult` , `PropDamage` and `TotalDamage` variables were recalculated. Recalculating the values for the entire dataset was not really necessary, but the code was much simpler.

```
sd$PROPDMGEXP[sd$REFNUM=='605943'] <- 'M'
sd$PropDamageMult <- convertExponentToMultiplier(sd$PROPDMGEXP)
sd$PropDamage  <- sd$PROPDMG * sd$PropDamageMult
sd$TotalDamage <- sd$PropDamage + sd$CropDamage
```

After checking the remaining top 5 by damage and health impact, it was found those are consistent with data available in the NOAA database. See appendix B for more information on the checks of the top individual events. 6. Given the poor consistency of the values found in the `EVTYPE` variable, it was decided to use the list of Event names from Section 2.1.1 of the Storm Data Documentation (https://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2Fpd01016005curr.pdf). Using various techniques, a new tidy variable named `EventType` was created containing one of these 48 event types or the value `UNCATEGORIZED` indicating the event was not included. All observations start off as `UNCATEGORIZED` and then updated with different approaches with one of the 48 values.

```
eventTypes <- c('Astronomical Low Tide', 'Avalanche', 'Blizzard', 'Coastal Flood', 'Cold/Wind Chill',
                'Debris Flow', 'Dense Fog', 'Dense Smoke', 'Drought', 'Dust Devil', 'Dust Storm',
                'Excessive Heat', 'Extreme Cold/Wind Chill', 'Flash Flood', 'Flood', 'Frost/Freeze',
                'Funnel Cloud', 'Freezing Fog', 'Hail', 'Heat', 'Heavy Rain', 'Heavy Snow', 'High Surf',
                'High Wind', 'Hurricane (Typhoon)', 'Ice Storm', 'Lake-Effect Snow', 'Lakeshore Flood',
                'Lightning', 'Marine Hail', 'Marine High Wind', 'Marine Strong Wind',
                'Marine Thunderstorm Wind', 'Rip Current', 'Seiche', 'Sleet', 'Storm Surge/Tide',
                'Strong Wind', 'Thunderstorm Wind', 'Tornado', 'Tropical Depression', 'Tropical Storm',
                'Tsunami', 'Volcanic Ash', 'Waterspout', 'Wildfire', 'Winter Storm', 'Winter Weather')
sd$EventType <- 'UNCATEGORIZED'  # start all EventTypes off as "Uncategorized"
```

The EVTYPE variable was first updated for consistency by removing all whitespace and making all upper case.

```
sd$EVTYPE <- toupper(trimws(sd$EVTYPE))
```

The inital pass of setting EventType values from EVTYPE data was a simple text matching approach based on: * Ignoring all whitespace and capitalization
* Ignoring all non-alpha characters * Allowing for plural variations (WIND/WIND**S**) * Allowing for verb variations (FLOOD/FLOOD**ING**)

```
regex <- "[^[:alpha:]]" # match all non-alpha
for(eventType in eventTypes) {
  strippedEventType <- toupper(gsub(regex, '', eventType))
  sd$EventType[gsub(regex, '', sd$EVTYPE) == strippedEventType] <- eventType
  sd$EventType[gsub(regex, '', sd$EVTYPE) == paste(strippedEventType, 'S',   sep='')] <- eventType
  sd$EventType[gsub(regex, '', sd$EVTYPE) == paste(strippedEventType, 'ING', sep='')] <- eventType
}
```

The next step of populating EventType was a manual mapping using EVTYPE values. Some were obvious abbreviations (TSTM WIND -> Thunderstorm Wind). Other values required reviewing the Storm Data Documentation (https://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2Fpd01016005curr.pdf) for better understanding. For example LANDSPOUT was mapped to Tornado and not Dust Devil because on page 75 it states:

> Landspouts and cold-air funnels, ultimately meeting the objective tornado criteria listed in Section 7.40.6, will be classified as Tornado events. This manual process was done iteratively while reviewing the damage and health impact totals for the remaining uncategorized EVTYPE values until it was determined that further work would not have any meaningful impact to the overall result of this report. See appendix C for the final EventType to EVTYPE value mappings and appendix D for more information on the EVTYPE values that were left uncategorized.

```
coastalFloodAliases <- c('ASTRONOMICAL HIGH TIDE', 'TIDAL FLOODING', 'COASTAL FLOODING/EROSION',
                         'COASTAL  FLOODING/EROSION', 'EROSION/CSTL FLOOD')
sd$EventType[sd$EVTYPE %in% coastalFloodAliases] <- 'Coastal Flood'

winterWeatherAliases <- c('LIGHT FREEZING RAIN', 'ICY ROADS', 'GLAZE', 'FREEZING RAIN',
                          'FREEZING DRIZZLE', 'LIGHT SNOW', 'LIGHT SNOWFALL', 'WINTER WEATHER/MIX',
                          'MIXED PRECIPITATION', 'MIXED PRECIP', 'WINTRY MIX', 'RAIN/SNOW',
                          'WINTER WEATHER MIX')
sd$EventType[sd$EVTYPE %in% winterWeatherAliases] <- 'Winter Weather'
heavySnowAliases <- c('EXCESSIVE SNOW', 'SNOW', 'HEAVY SNOW SHOWER', 'SNOW SQUALL', 'SNOW SQUALLS')
sd$EventType[sd$EVTYPE %in% heavySnowAliases] <- 'Heavy Snow'
highWindAliases <- c('WIND', 'WINDS', 'GUSTY WINDS', 'GUSTY WIND', 'HIGH WIND (G40)',
                     'NON TSTM WIND',  'NON-TSTM WIND', 'WIND DAMAGE', 'NON TSTM WIND',
                     'NON-SEVERE WIND DAMAGE', 'GRADIENT WIND')
sd$EventType[sd$EVTYPE %in% highWindAliases] <- 'High Wind'
freezeAliases <- c('FREEZE', 'DAMAGING FREEZE', 'EARLY FROST', 'FROST', 'AGRICULTURAL FREEZE',
                   'HARD FREEZE', 'UNSEASONABLY COLD', 'UNSEASONABLE COLD')
sd$EventType[sd$EVTYPE %in% freezeAliases] <- 'Frost/Freeze'
extremeColdAliases <- c('EXTREME WINDCHILL', 'EXTREME COLD')
sd$EventType[sd$EVTYPE %in% extremeColdAliases] <- 'Extreme Cold/Wind Chill'
floodAliases <- c('RIVER FLOODING', 'RIVER FLOOD', 'URBAN/SML STREAM FLD', 'URBAN FLOOD')
sd$EventType[sd$EVTYPE %in% floodAliases] <- 'Flood'
flashFloodAliases <- c('FLASH FLOOD/FLOOD', 'FLOOD/FLASH/FLOOD')
sd$EventType[sd$EVTYPE %in% flashFloodAliases] <- 'Flash Flood'
thunderstormAliases <- c('TSTM WIND', 'TSTM WINDS', 'THUNDERSTORM', 'THUNDERSTORMS',
                         'THUNDERSTORM WINDSS', 'THUNDERSTORMS WINDS', 'DRY MICROBURST',
                         'TSTM WIND (G40)', 'THUNDERSTORM WIND/ TREES', 'MICROBURST',
                         'WET MICROBURST', 'THUNDERTORM WINDS', 'THUNDERSTORMS WIND',
                         'SEVERE THUNDERSTORM WINDS', 'TSTM WIND 55', 'THUNDERSTORM WIND 60 MPH',
                         'TSTM WIND (G45)', 'SEVERE THUNDERSTORM', 'THUDERSTORM WINDS',
                         'THUNDEERSTORM WINDS', 'THUNDERESTORM WINDS', 'TSTM WIND 40',
                         'TSTM WIND G45', 'TSTM WIND  (G45)', 'TSTM WIND (41)', 'TSTM WIND 45',
                         'TSTM WIND (G35)', 'TSTM WIND AND LIGHTNING', 'TSTM WIND/HAIL',
                         'THUNDERSTORM WIND (G40)')
sd$EventType[sd$EVTYPE %in% thunderstormAliases] <- 'Thunderstorm Wind'
hailAliases <- c('HAIL DAMAGE', 'SMALL HAIL', 'HAILSTORM')
sd$EventType[sd$EVTYPE %in% hailAliases] <- 'Hail'
hurricaneAliases <- c('HURRICANE', 'TYPHOON', 'HURRICANE OPAL', 'HURRICANE ERIN',
                      'HURRICANE EDOUARD', 'HURRICANE EMILY', 'HURRICANE FELIX',
                      'HURRICANE GORDON', 'HURRICANE OPAL/HIGH WINDS')
sd$EventType[sd$EVTYPE %in% hurricaneAliases] <- 'Hurricane (Typhoon)'
highSurfAliases <- c('HEAVY SURF/HIGH SURF', 'HEAVY SURF', 'HIGH SURF ADVISORY')
sd$EventType[sd$EVTYPE %in% highSurfAliases] <- 'High Surf'
wildfireAliases = c('WILD/FOREST FIRE', 'BRUSH FIRE')
sd$EventType[sd$EVTYPE %in% wildfireAliases] <- 'Wildfire'
heatAliases = c('UNSEASONABLY WARM', 'WARM WEATHER')
sd$EventType[sd$EVTYPE %in% heatAliases] <- 'Heat'
excessiveHeatAliases = c('HEAT WAVE', 'RECORD HEAT')
sd$EventType[sd$EVTYPE %in% excessiveHeatAliases] <- 'Excessive Heat'
heavyRainAliases = c('TORRENTIAL RAINFALL', 'RAIN', 'UNSEASONAL RAIN')
sd$EventType[sd$EVTYPE %in% heavyRainAliases]  <- 'Heavy Rain'
# one-offs
sd$EventType[sd$EVTYPE == 'LANDSPOUT']         <- 'Tornado'
sd$EventType[sd$EVTYPE == 'FOG']               <- 'Dense Fog'
sd$EventType[sd$EVTYPE == 'MARINE TSTM WIND']  <- 'Marine Thunderstorm Wind'
sd$EventType[sd$EVTYPE == 'LANDSLIDE']         <- 'Debris Flow'
sd$EventType[sd$EVTYPE == 'STORM SURGE']       <- 'Storm Surge/Tide'
sd$EventType[sd$EVTYPE == 'COLD']              <- 'Cold/Wind Chill'
```

Results

Event Types Most Harmful to Population Health

```
top_health <- head(
  arrange(
    aggregate(
      cbind(FATALITIES, INJURIES, PopulationHealthImpact) ~ EventType, sd, FUN = sum),
    desc(PopulationHealthImpact)
  ),
  n=5
)
kable(
  top_health,
  caption = 'Top 5 Event Types Most Harmful to Population Health'
)
```

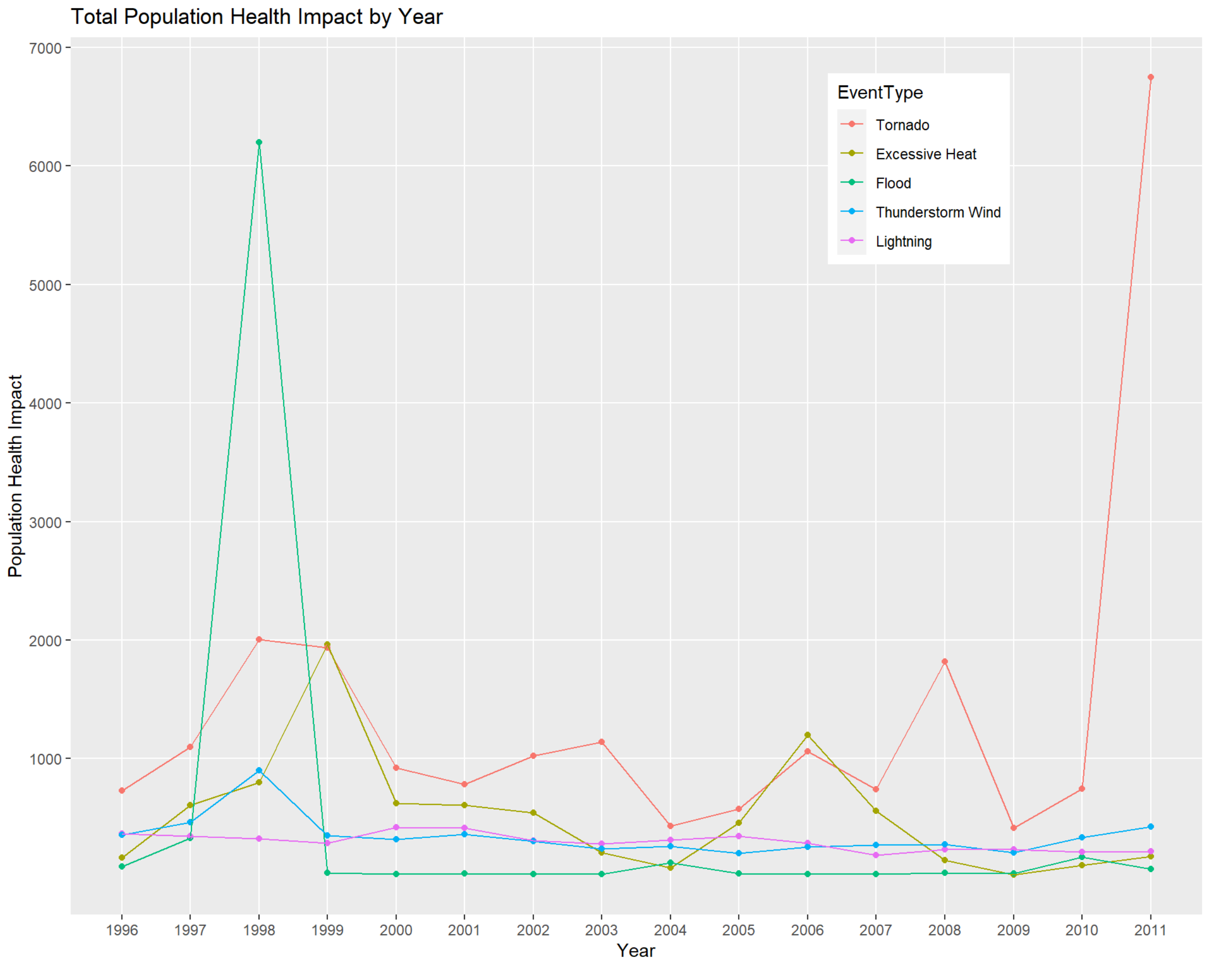
Top 5 Event Types Most Harmful to Population Health

EventType	FATALITIES	INJURIES	PopulationHealthImpact
Tornado	1511	20667	22178
Excessive Heat	1799	6461	8260
Flood	444	6838	7282
Thunderstorm Wind	382	5154	5536
Lightning	651	4141	4792

Tornado events top the list here, with over two and half times the health impact of second place, which is Excessive Heat. Excessive Heat is worth noting however due to the fact that even though it is far behind tornados in total health impact, but has the most fatalities overall.

Next we will look a bit deeper at the data, plotting the yearly total health impact for these top 5.

```
health_by_type_and_year <- aggregate(
  cbind(FATALITIES, INJURIES, PopulationHealthImpact) ~ EventType + year(BeginDate),
  sd,
  FUN=sum
)
names(health_by_type_and_year) <- c('EventType', 'Year', 'Fatalities', 'Injuries', 'PopulationHealthImpact')
health_by_type_and_year <- health_by_type_and_year[health_by_type_and_year$EventType %in% top_health$EventType,]
health_by_type_and_year$EventType <- with(health_by_type_and_year, reorder(EventType, -PopulationHealthImpact))
ggplot(health_by_type_and_year, aes(x=Year, y=PopulationHealthImpact, colour = EventType)) +
  geom_point() + geom_line() +
  scale_x_continuous(breaks = unique(health_by_type_and_year$Year)) +
  scale_y_continuous(
    'Population Health Impact',
    breaks = seq(1000, 7000, by=1000)
  ) +
  ggtitle("Total Population Health Impact by Year") +
  theme(
    legend.position = c(0.75, 0.85),
    panel.grid.minor = element_blank()
  )
```



Here we see two years with significant outliers. In 1998 there was an extremely high health related impact due to flood events. Looking at the NOAA Summary of Natural Hazard Statistics for 1998 (http://www.nws.noaa.gov/om/hazstats/sum98.pdf) shows that a flood in south-central Texas caused over 6,000 injuries accounting for most of that year’s total. The Tornado spike in 2011 can be accounted for due to record breaking spring and summer tornado season according to the NOAA Tornado Annual 2011 Report (https://www.ncdc.noaa.gov/sotc/tornadoes/201113).

Looking at the plot, Tornados have a solid yearly trend despite the record breaking year, so their number one position is not due to that year alone. Flood events however have an overall low yearly trend in comparison to the other top 5 except for 1998. Without this year, flood events would have been in last place instead of third amongst the current top 5. Additional analysis would be needed, but there is good chance it would not have even made the top 5 at all without the 1998 Texas floods.

Event Types with Greatest Economic Consequences

```
top_damage <- head(
  arrange(
    aggregate(
      cbind(CropDamage, PropDamage, TotalDamage) ~ EventType, sd, FUN=sum),
    desc(TotalDamage)
  ),
  n=5
)
kable(
  top_damage,
  format.args = list(big.mark = ","),
  caption = 'Top 5 Event Types with Greatest Economic Consequences'
)
```

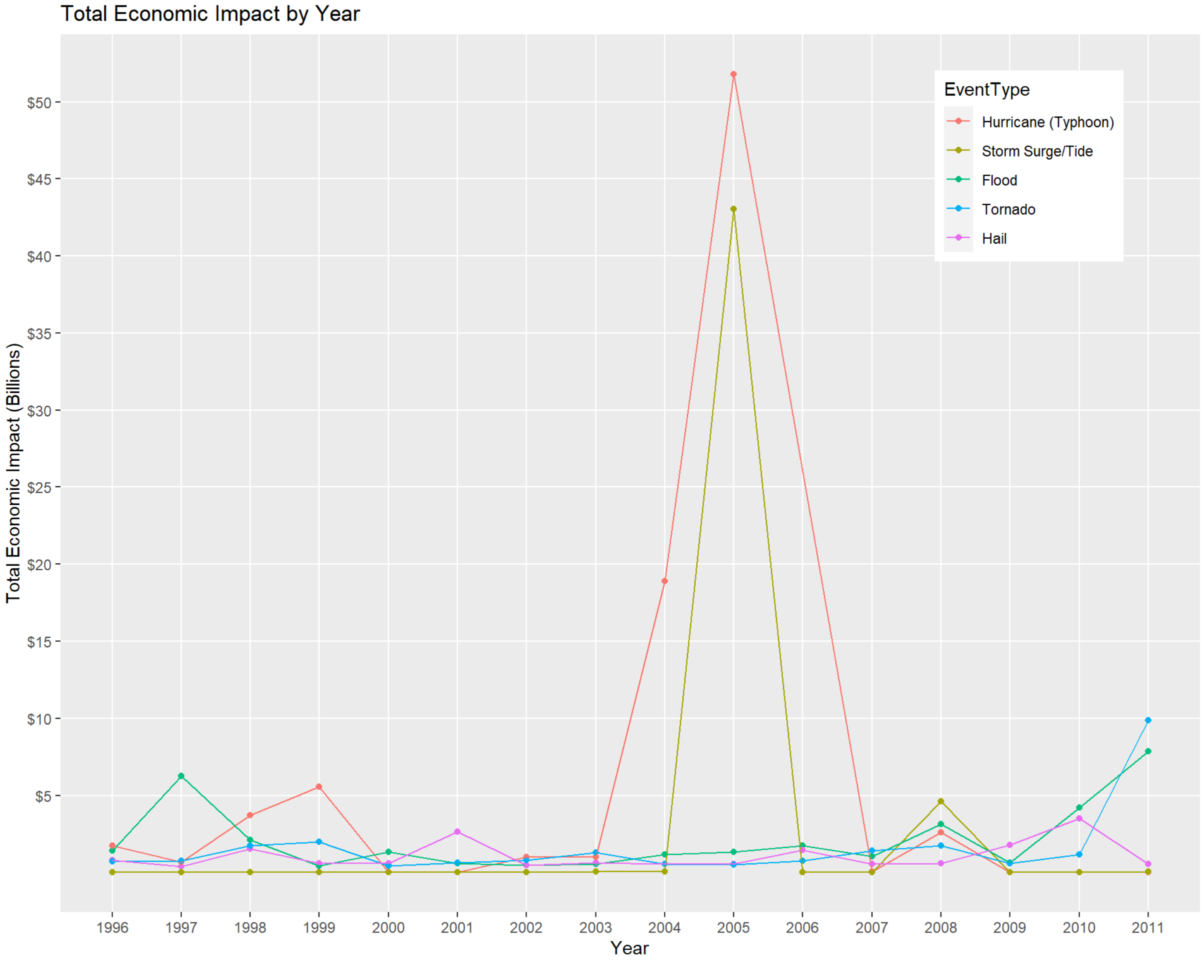
Top 5 Event Types with Greatest Economic Consequences

EventType	CropDamage	PropDamage	TotalDamage
Hurricane (Typhoon)	5,350,107,800	81,718,889,010	87,068,996,810
Storm Surge/Tide	855,000	47,834,724,000	47,835,579,000
Flood	5,013,161,500	29,244,580,200	34,257,741,700
Tornado	283,425,010	24,616,952,710	24,900,377,720
Hail	2,496,822,450	14,595,213,420	17,092,035,870

Here we see that Hurricane (Typhoon) events top the list with $87 billion, which is almost double the next in line which is Storm Surge/Tide events at $47 billion. One interesting note is that Flood events caused almost as much crop damage as hurricanes despite being a distant third place overall.

Again, we will look at the yearly trend for these top five.

```
damage_by_type_and_year <-  aggregate(
  cbind(CropDamage,PropDamage,TotalDamage)~EventType+year(BeginDate),
  sd,
  FUN=sum
)
names(damage_by_type_and_year) <- c('EventType', 'Year', 'CropDamage', 'PropDamage', 'TotalDamage')
sd_dmg_yearly <- damage_by_type_and_year[damage_by_type_and_year$EventType %in% top_damage$EventType,]
sd_dmg_yearly$EventType <- with(sd_dmg_yearly, reorder(EventType, -TotalDamage))
ggplot(sd_dmg_yearly, aes(Year, TotalDamage / 10^9, colour = EventType)) +
  geom_point() + geom_line() +
  scale_x_continuous(breaks = unique(damage_by_type_and_year$Year)) +
  scale_y_continuous(
    'Total Economic Impact (Billions)',
    labels = scales::dollar,
    breaks = seq(5,50, by=5)
  ) +
  ggtitle('Total Economic Impact by Year') +
  theme(
    legend.position = c(0.85, 0.85),
    panel.grid.minor = element_blank()
  )
```



Like in the previous yearly trend, we see a couple of significant outliers, but this time they both occur in the same year of 2005 with Hurricane (Typhoon) and Storm Surge/Tide events. The significant Hurricane event for 2005 was Hurricane Katrina according to the NOAA 2005 Summary of Natural Hazard Statistics (http://www.nws.noaa.gov/om/hazstats/sum05.pdf) where it is noted that Katrina had an estimated $93 billion in claims. While Storm Surge/Tide events are not called out in the NOAA summary, the $93 billion seems to correlate with the combined values of Hurricanes and Storm surges for that year.

Similarly, as with the top five events for population health impact, the top five event list might look different if it were not for this year with the significant outliers. Additional analysis would be needed, but hurricane event’s number one position could be in jeopardy without 2005 and storm surge might not have even made the list at all without it.