# Storm Event Types Impacting Population Health and Economy

JLei

4/22/2020

### Synopsis

Storms and other severe weather events can impact population health and have economic consequences for communities and cities. Identifying the major evernts that result in fatalities, injuries, and property damages can help drive public policies in preventing such outcomes. The analysis descibed in this article explored the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database that covered events between the year 1950 and end in November 2011. From these data, we found that tornado is the most harmful with respect to population health according to the total incident numbers of fatalities and injuries it caused. Tornado also has the greatest economic consequence with the property damage it caused, while hail has the greatest economic consequence with the damage on crops.

### Data Processing

The data used for the analysis come in the form of a comma-separated-value (CSV) file compressed via the bzip2 algorithm to reduce its size. Below lists more information:

- Storm Data (47Mb)
- National Weather Service Storm Data Documentation
- National Climatic Data Center Storm Events FAQs

### Load Data

First, create a directory /data in the set directory to download the dataset to. The file is in .csv.bz2 format, and can be read in by calling *read.csv()* function.

```
devtools::install_github('yihui/tinytex')
```

```
## Downloading GitHub repo yihui/tinytex@master

## xfun (0.12 -> 0.13) [CRAN]

## Installing 1 packages: xfun

##
## The downloaded binary packages are in
##  /var/folders/8t/nhjx2nsn237d78tgxk_7nk0h0000gn/T//RtmpaArN5w/downloaded_packages
##      checking for file '/private/var/folders/8t/nhjx2nsn237d78tgxk_7nk0h0000gn/T/RtmpaArN5w/remotes5
##    -  preparing 'tinytex':
##      checking DESCRIPTION meta-information ...  v  checking DESCRIPTION meta-information
##    -  checking for LF line-endings in source and make files and shell scripts
##    -  checking for empty or unneeded directories
##    -  building 'tinytex_0.22.2.tar.gz'
##
##
```

```r
options(tinytex.verbose = TRUE)
## Set to the desired directory and create a data directory
if (!file.exists("data")) {
    dir.create("data")
}

## Download the data
fileUrl <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
download.file(fileUrl, destfile = "./data/stormdata.csv.bz2", method = "curl")

## Read in the csv data
storm <- read.csv("./data/stormdata.csv.bz2")
```

Let's check the structure, dimension, column names, and first few rows of the data **storm**.

```r
str(storm)
```

```
## '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 
##  $ 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 
##  $ 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 ""," Christiansburg",..: 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 ""," CANTON"," TULIA",..: 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","%SD",..: 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 "","\t","\t\t",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ REFNUM    : num  1 2 3 4 5 6 7 8 9 10 ...
```

```r
dim(storm)
```

```
## [1] 902297     37
```

```r
names(storm)
```

```
##  [1] "STATE__"    "BGN_DATE"   "BGN_TIME"   "TIME_ZONE"  "COUNTY"
##  [6] "COUNTYNAME" "STATE"      "EVTYPE"     "BGN_RANGE"  "BGN_AZI"
## [11] "BGN_LOCATI" "END_DATE"   "END_TIME"   "COUNTY_END" "COUNTYENDN"
## [16] "END_RANGE"  "END_AZI"    "END_LOCATI" "LENGTH"     "WIDTH"
## [21] "F"          "MAG"        "FATALITIES" "INJURIES"   "PROPDMG"
## [26] "PROPDMGEXP" "CROPDMG"    "CROPDMGEXP" "WFO"        "STATEOFFIC"
## [31] "ZONENAMES"  "LATITUDE"   "LONGITUDE"  "LATITUDE_E" "LONGITUDE_"
## [36] "REMARKS"    "REFNUM"
```

```r
head(storm, 3)
```

```
##   STATE__           BGN_DATE BGN_TIME TIME_ZONE COUNTY COUNTYNAME STATE  EVTYPE
## 1       1 4/18/1950 0:00:00     0130       CST     97     MOBILE    AL TORNADO
## 2       1 4/18/1950 0:00:00     0145       CST      3    BALDWIN    AL TORNADO
## 3       1 2/20/1951 0:00:00     1600       CST     57    FAYETTE    AL TORNADO
##   BGN_RANGE BGN_AZI BGN_LOCATI END_DATE END_TIME COUNTY_END COUNTYENDN
## 1         0                                               0         NA
## 2         0                                               0         NA
## 3         0                                               0         NA
##   END_RANGE END_AZI END_LOCATI LENGTH WIDTH F MAG FATALITIES INJURIES PROPDMG
## 1         0                      14.0   100 3   0          0       15    25.0
## 2         0                       2.0   150 2   0          0        0     2.5
## 3         0                       0.1   123 2   0          0        2    25.0
##   PROPDMGEXP CROPDMG CROPDMGEXP WFO STATEOFFIC ZONENAMES LATITUDE LONGITUDE
## 1          K       0                                         3040      8812
## 2          K       0                                         3042      8755
## 3          K       0                                         3340      8742
##   LATITUDE_E LONGITUDE_ REMARKS REFNUM
## 1       3051       8806              1
## 2          0          0              2
## 3          0          0              3
```

The weather event types are tabulated in the column with variable EVTYPE. Population health outcomes are indicated by both FATALITIES and INJURIES variables. Economic consequences are indicated by property damage PROPDMG and crops damage CROPDMG variables.

Missing values are a common problem with environmental and weather related data, so we can check to see what portion of recorded data are missing (coded as NA).

```r
mean(is.na(storm$FATALITIES))
```

```
## [1] 0
```

```r
mean(is.na(storm$INJURIES))
```

```
## [1] 0
```

```r
mean(is.na(storm$PROPDMG))
```

```
## [1] 0
```

```r
mean(is.na(storm$CROPDMG))
```

```
## [1] 0
```

There is no missing value for the variables of interest.

For population health related variables, let's total up the incident numbers for each variable, categorized by event type. The package *dplyr* can help process the data this way.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
pophlth <- storm %>%
        group_by(EVTYPE) %>%
        summarise(Fatalities = sum(FATALITIES),
                  Injuries = sum(INJURIES))
```

```
summary(pophlth)
```

```
##                    EVTYPE      Fatalities          Injuries
##     HIGH SURF ADVISORY:  1   Min.   :   0.00   Min.   :    0.0
##    COASTAL FLOOD       :  1   1st Qu.:   0.00   1st Qu.:    0.0
##    FLASH FLOOD         :  1   Median :   0.00   Median :    0.0
##    LIGHTNING           :  1   Mean   :  15.38   Mean   :  142.7
##    TSTM WIND           :  1   3rd Qu.:   0.00   3rd Qu.:    0.0
##    TSTM WIND (G45)     :  1   Max.   :5633.00   Max.   :91346.0
##   (Other)              :979
```

The summary statitstics of Fatalities sum and the Injuries sum shows that a lot of events account for 0 incidents, so next we'll re-summarize the pophlth to exclude the events that never caused incidents.

```
pophlth <- pophlth %>%
        filter(Fatalities > 0 | Injuries > 0)
summary(pophlth)
```

```
##          EVTYPE      Fatalities          Injuries
##  AVALANCE    :  1   Min.   :   0.00   Min.   :    0.00
##  AVALANCHE   :  1   1st Qu.:   1.00   1st Qu.:    0.00
##  BLACK ICE   :  1   Median :   2.00   Median :    2.00
##  BLIZZARD    :  1   Mean   :  68.84   Mean   :  638.76
##  blowing snow:  1   3rd Qu.:  10.25   3rd Qu.:   35.25
##  BLOWING SNOW:  1   Max.   :5633.00   Max.   :91346.00
##  (Other)     :214
```

For economic consequences related variables, let's total up the financial costs (in k) for each variable, not accounting for the $0 damage and categorized by event type.

```
econ <- storm %>%
        filter(PROPDMG > 0 | CROPDMG > 0) %>%
        group_by(EVTYPE) %>%
        summarise(Property = sum(PROPDMG),
                  Crop = sum(CROPDMG))
summary(econ)
```

```
##                     EVTYPE       Property            Crop
##     HIGH SURF ADVISORY:  1   Min.    :       0   Min.    :     0.0
##    FLASH FLOOD         :  1   1st Qu.:       5   1st Qu.:     0.0
##    TSTM WIND           :  1   Median :      50   Median :     0.0
##    TSTM WIND (G45)     :  1   Mean   :   25254   Mean   :  3196.8
##   ?                    :  1   3rd Qu.:     501   3rd Qu.:     8.6
##   AGRICULTURAL FREEZE  :  1   Max.   :3212258   Max.   :579596.3
##   (Other)              :425
```

It is clear that there are outliers (max totals) for Fatalities, Injuries, Property damage, and Crop damage, and let's next identify the event type that caused these max totals for each variable.

```r
pophlth[which.max(pophlth$Fatalities), 1:2]
```

```
## # A tibble: 1 x 2
##   EVTYPE  Fatalities
##   <fct>        <dbl>
## 1 TORNADO       5633
```

```r
pophlth[which.max(pophlth$Injuries), c(1,3)]
```

```
## # A tibble: 1 x 2
##   EVTYPE  Injuries
##   <fct>      <dbl>
## 1 TORNADO    91346
```

```r
econ[which.max(econ$Property), 1:2]
```

```
## # A tibble: 1 x 2
##   EVTYPE  Property
##   <fct>      <dbl>
## 1 TORNADO 3212258.
```

```r
econ[which.max(econ$Crop), c(1,3)]
```

```
## # A tibble: 1 x 2
##   EVTYPE    Crop
##   <fct>    <dbl>
## 1 HAIL   579596.
```

TORNADO is the weather event type that is most harmful with respect to population health, and it has the greatest economic consequences with Property Damage, while HAIL has the greatest economic consequences when we totaled up the incidents and financial costs over the years 1950-2011.

To look at the impact of the two identified weather events on population health and economic burden, let's further look at the total incident numbers and financial costs, respectively, over the years of 1950-2011.

Now we need to massage the BGN_DATE variable into a date format, and extract the year to a new column under the variable name *year*.

```r
storm$year <- substring(as.Date(storm$BGN_DATE, format="%m/%d/%Y"), 1, 4)
```

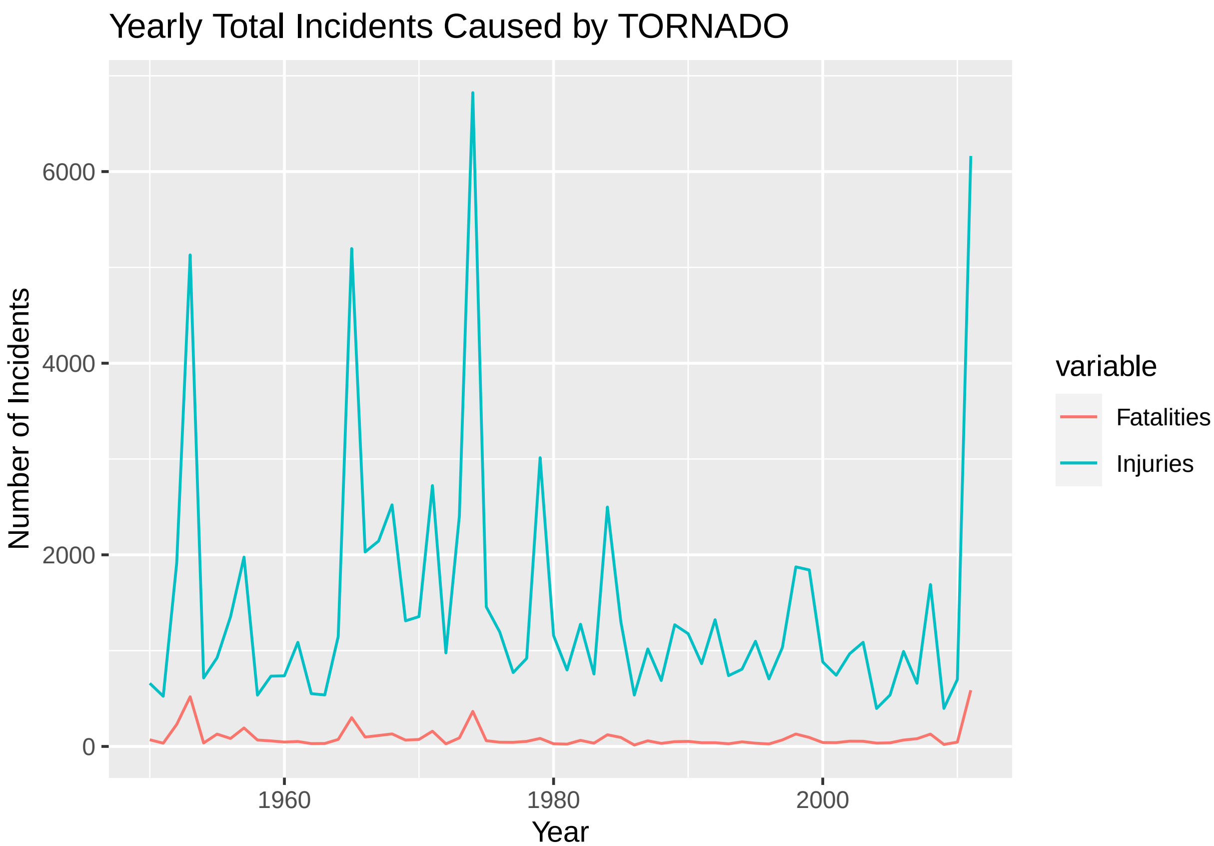
### Results

The package *ggplot()* is utilized to plot the progress for each variable.

```r
## subset the TORNADO data, arrange it by year, and total up the incident numbers
tornado <- storm %>%
        filter(EVTYPE == "TORNADO") %>%
        mutate(year = as.numeric(year)) %>%
```

```
        group_by(year) %>%
        summarise(Fatalities = sum(FATALITIES),
                  Injuries = sum(INJURIES))

## load ggplot2 package
library(ggplot2)
## plot the yearly progression of incidents caused by TORNADO
t <- ggplot(tornado, aes(year, y=value, color=variable))
t + geom_line(aes(y = Fatalities, col = "Fatalities")) +
        geom_line(aes(y = Injuries, col = "Injuries")) +
        labs(title = "Yearly Total Incidents Caused by TORNADO",
             x = "Year",
             y = "Number of Incidents")
```

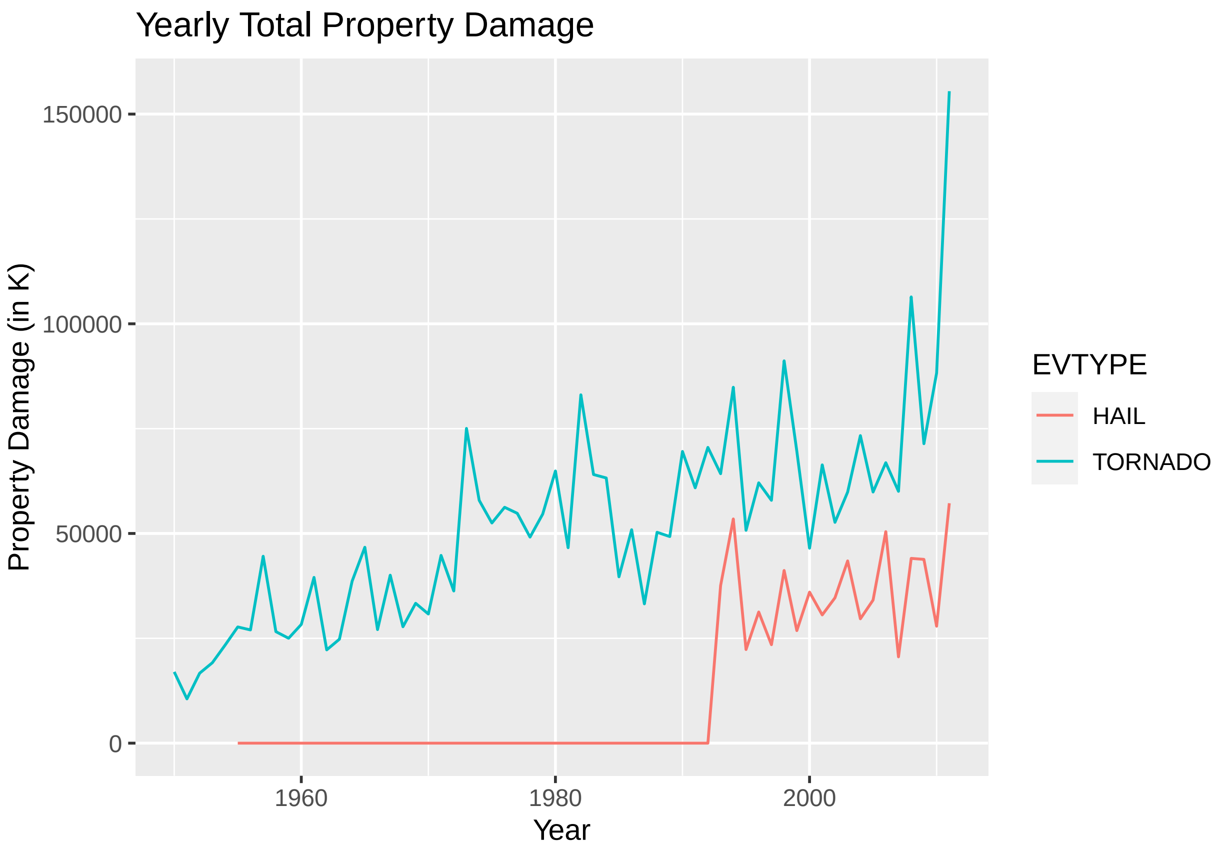



Historically, there were 3 or 4 times that incidents of Fatalities and Injuries shot up significantly (referring to the spikes), this shows how drastically tornado impacts community safety.

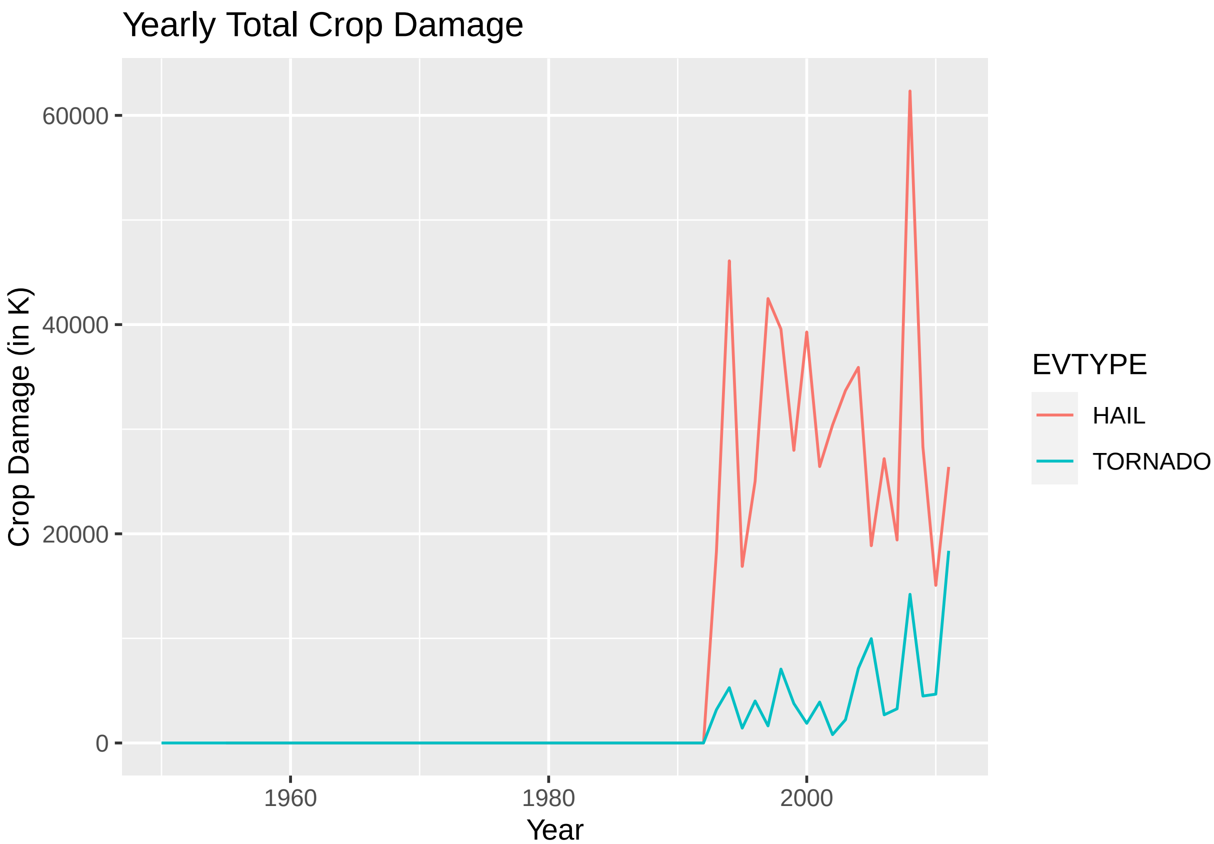
```
## subset the HAIL data, arrange it by year, and total up the financial costs
tor_hail_prop <- storm %>%
        filter(EVTYPE == "HAIL" | EVTYPE == "TORNADO") %>%
        mutate(year = as.numeric(year)) %>%
        group_by(EVTYPE, year) %>%
        summarise(Property = sum(PROPDMG))
tor_hail_crop <- storm %>%
        filter(EVTYPE == "HAIL" | EVTYPE == "TORNADO") %>%
        mutate(year = as.numeric(year)) %>%
        group_by(EVTYPE, year) %>%
```

```
        summarise(Crop = sum(CROPDMG))

## plot the yearly progression of property damange, comparing TORNADO and HAIL
thp <- ggplot(tor_hail_prop, aes(year, Property, color=EVTYPE))
thp + geom_line() +
        labs(title = "Yearly Total Property Damage",
             x = "Year",
             y = "Property Damage (in K)")
```




```
## plot the yearly progression of crop damange, comparing TORNADO and HAIL
thc <- ggplot(tor_hail_crop, aes(year, Crop, color=EVTYPE))
thc + geom_line() +
        labs(title = "Yearly Total Crop Damage",
             x = "Year",
             y = "Crop Damage (in K)")
```




Tornado not only impacts population health and community safety, but it also destroys properties which in turn causes financial losses. The plot "Yearly Total Property Damage" shows an overall increase trend in property damage over of the year by Tornado.

There was no damage recorded for hail until mid 1950's. With the available recordings, Property and Crop damage caused by hail did not pick up huge financial values until 1992-1993, and also followed a increasing trend. The yearly total damage on both property and crop was at the scale of millions of dollar, showing the devastating economic consequences.