Storm Data Analysis

Suzannah Herron

2/4/2022

Synopsis

This data comes from the US National Oceanic and Atmospheric Administration's (NOAA) storm database. There was a lot of cleaning that needed to be done, especially regarding the names of the event types.
The goal is to look at:
1. Which weather event types are the deadliest (first plot)
2. Which weather event types have the greatest economic impact (second plot)
I also decided to find which states were the safest and most dangerous to live in (third plot).

Data Processing

Checking to see if dataset exists already. If not, then download and unzip.

```
file_name <- "storms.csv.bz2"
if (!file.exists(file_name)) {
  url <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
  download.file(url, file_name, method = "curl")
}
if (!file.exists(file_name)) {
  unzip(file_name)
}
storms <- read.csv(file_name)
str(storms)
```

```
## 'data.frame':    902297 obs. of  37 variables:
##  $ STATE__   : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ BGN_DATE  : chr  "4/18/1950 0:00:00" "4/18/1950 0:00:00" "2/20/1951 0:00:00" "6/8/1951 0:00:00" .
##  $ BGN_TIME  : chr  "0130" "0145" "1600" "0900" ...
##  $ TIME_ZONE : chr  "CST" "CST" "CST" "CST" ...
##  $ COUNTY    : num  97 3 57 89 43 77 9 123 125 57 ...
##  $ COUNTYNAME: chr  "MOBILE" "BALDWIN" "FAYETTE" "MADISON" ...
##  $ STATE     : chr  "AL" "AL" "AL" "AL" ...
##  $ EVTYPE    : chr  "TORNADO" "TORNADO" "TORNADO" "TORNADO" ...
##  $ BGN_RANGE : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BGN_AZI   : chr  "" "" "" "" ...
##  $ BGN_LOCATI: chr  "" "" "" "" ...
##  $ END_DATE  : chr  "" "" "" "" ...
##  $ END_TIME  : chr  "" "" "" "" ...
##  $ 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   : chr  "" "" "" "" ...
##  $ END_LOCATI: chr  "" "" "" "" ...
##  $ 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: chr  "K" "K" "K" "K" ...
##  $ CROPDMG   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CROPDMGEXP: chr  "" "" "" "" ...
##  $ WFO       : chr  "" "" "" "" ...
##  $ STATEOFFIC: chr  "" "" "" "" ...
##  $ ZONENAMES : chr  "" "" "" "" ...
##  $ 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   : chr  "" "" "" "" ...
##  $ REFNUM    : num  1 2 3 4 5 6 7 8 9 10 ...
```

Format `BGN_DATE` to be a date then add a column with the year to use for factoring later.

```
library(lubridate)
library(dplyr)
library(ggplot2)
```

```
storms$BGN_DATE <- mdy_hms(storms$BGN_DATE)
storms$YEAR <- year(storms$BGN_DATE)
# str(storms)
```

Cutting down the dataset to relevant columns/variables for analysis for bodily, property, and crop damage.

Variables

STATE: State abbreviation (I am keeping this in case I want to try mapping)
EVTYPE: Type of event
FATALITIES: Number of fatalities caused by event
INJURIES: Number of injuries caused by event
PROPDMG: USD amount of property damage
PROPDMGEXP: Multiplying factor {exponent} for PROPDMG
CROPDMG: USD amount of crop damage
CROPDMGEXP: Multiplying factor (exponent) for CROPDMG
YEAR: Starting year of event extracted from `storms$BGN_DATE`

```
cols <- c( "STATE", "EVTYPE", "FATALITIES", "INJURIES", "PROPDMG", "PROPDMGEXP", "CROPDMG", "CROPDMGEXP
storms1 <- storms[ , cols]
# str(storms1)
```

Since we are looking at the different effects of each event type, a creation of a dataframe grouped by event types is necessary.
Checking the types of events:

```
types <- unique(storms1$EVTYPE)
str(types)
```

```
##  chr [1:985] "TORNADO" "TSTM WIND" "HAIL" "FREEZING RAIN" "SNOW" ...
```

Some of the event types are all upper case and some are a combo of upper and lower case. Changed all to lowercase and the number of events went from 985 to 898

```
storms2 <- storms1
storms2$EVTYPE <- tolower(storms2$EVTYPE)
types <- unique(storms2$EVTYPE)
str(types)
```

```
##  chr [1:898] "tornado" "tstm wind" "hail" "freezing rain" "snow" ...
```

There are several event names with whitespace in the beginning so that was trimmed and the number of events is 890

```
storms2$EVTYPE <- trimws(storms2$EVTYPE, which = "both")
types <- unique(storms2$EVTYPE)
str(types)
```

```
##  chr [1:890] "tornado" "tstm wind" "hail" "freezing rain" "snow" ...
```

The next step would be to group the event types into larger categories to make it easier to analyze.
Categories will be:
tornado, wind, hail, rain, flood, hurricane, lightning, cold (snow, ice, blizzards, winter storms), fire
Afterwards, I will check the events categorized as "other" and reassess.

```
storms3 <- storms2
storms3$CAT <- "other"
storms3$CAT[grep("tornado", storms3$EVTYPE)] <- "tornado"
storms3$CAT[grep("wind", storms3$EVTYPE)] <- "wind"
storms3$CAT[grep("hail", storms3$EVTYPE)] <- "hail"
storms3$CAT[grep("rain", storms3$EVTYPE)] <- "rain"
storms3$CAT[grep("flood", storms3$EVTYPE)] <- "flood"
storms3$CAT[grep("hurricane", storms3$EVTYPE)] <- "hurricane"
storms3$CAT[grep("lightning", storms3$EVTYPE)] <- "thunderstorm"
storms3$CAT[grep("snow", storms3$EVTYPE)] <- "cold"
storms3$CAT[grep("ice", storms3$EVTYPE)] <- "cold"
storms3$CAT[grep("blizzard", storms3$EVTYPE)] <- "cold"
storms3$CAT[grep("winter", storms3$EVTYPE)] <- "cold"
storms3$CAT[grep("fire", storms3$EVTYPE)] <- "fire"
sort(table(storms3$CAT))
```

```
## 
##    hurricane         fire         rain thunderstorm        other         cold 
##          288         4240        12168        15775        30167        42155 
##      tornado        flood         hail         wind 
##        60699        82713       290400       363692
```

Take a look what is still in the "other" category:

```
other <- storms3
other <- storms3[storms3$CAT == "other", ]
other <- other[ , c(2,10)]
types_other <- unique(other$EVTYPE)
head(types_other)
```

```
## [1] "record cold"       "dense fog"         "rip current"
## [4] "thunderstorm wins" "funnel cloud"      "heat"
```

Categorize more and changing "fire" to "heat" and changing "lightning" to "thunderstorm". Adding categories "volcano" and "dry". (I changed some of the code in chunk 9.)

```
storms3$CAT[grep("cold", storms3$EVTYPE)] <- "cold"
storms3$CAT[grep("record low", storms3$EVTYPE)] <- "cold"
storms3$CAT[grep("fire", storms3$EVTYPE)] <- "heat"
storms3$CAT[grep("heat", storms3$EVTYPE)] <- "heat"
storms3$CAT[grep("hot", storms3$EVTYPE)] <- "heat"
storms3$CAT[grep("warm", storms3$EVTYPE)] <- "heat"
storms3$CAT[grep("record high", storms3$EVTYPE)] <- "cold"
storms3$CAT[grep("icy", storms3$EVTYPE)] <- "cold"
storms3$CAT[grep("freez", storms3$EVTYPE)] <- "cold"
storms3$CAT[grep("sleet", storms3$EVTYPE)] <- "cold"
storms3$CAT[grep("showers", storms3$EVTYPE)] <- "rain"
storms3$CAT[grep("precipitation", storms3$EVTYPE)] <- "cold"
storms3$CAT[grep("thunderstorm", storms3$EVTYPE)] <- "thunderstorm"
storms3$CAT[grep("microburst", storms3$EVTYPE)] <- "thunderstorm"
storms3$CAT[grep("tstm", storms3$EVTYPE)] <- "thunderstorm"
storms3$CAT[grep("volcan", storms3$EVTYPE)] <- "volcano"
storms3$CAT[grep("dry", storms3$EVTYPE)] <- "dry"
storms3$CAT[grep("driest", storms3$EVTYPE)] <- "dry"
storms3$CAT[grep("drought", storms3$EVTYPE)] <- "dry"
```

```
other <- storms3
other <- storms3[storms3$CAT == "other", ]
other <- other[ , c(2,10)]
types_other <- unique(other$EVTYPE)
length(types_other)
```

```
## [1] 222
```

```
length(unique(storms3$CAT))
```

```
## [1] 12
```

Now I have 12 categories with 222 event types categorized as "other". Just as a side note, looking through the event types, there are a lot of typos in this data.

Splitting Data

Create two different dataframes with injuries/fatalities and financial impact.

```
fatal_inj <- storms3[ , c(1:4, 9, 10)]
head(fatal_inj)
```

```
##   STATE  EVTYPE FATALITIES INJURIES YEAR     CAT
## 1    AL tornado          0       15 1950 tornado
## 2    AL tornado          0        0 1950 tornado
## 3    AL tornado          0        2 1951 tornado
## 4    AL tornado          0        2 1951 tornado
## 5    AL tornado          0        2 1951 tornado
## 6    AL tornado          0        6 1951 tornado
```

```
prop_crop <- storms3[ , c(1, 2, 5:10)]
head(prop_crop)
```

```
##   STATE  EVTYPE PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP YEAR     CAT
## 1    AL tornado    25.0          K       0            1950 tornado
## 2    AL tornado     2.5          K       0            1950 tornado
## 3    AL tornado    25.0          K       0            1951 tornado
## 4    AL tornado     2.5          K       0            1951 tornado
## 5    AL tornado     2.5          K       0            1951 tornado
## 6    AL tornado     2.5          K       0            1951 tornado
```

Analysis

Which types of events are most harmful to population health?

The first step is to find the totals for injuries and fatalities for each event type per year.

```
sum_fatal <- aggregate(FATALITIES ~ CAT + YEAR, fatal_inj, sum)
sum_inj <- aggregate(INJURIES ~ CAT + YEAR, fatal_inj, sum)
sum_fi <- cbind(sum_fatal, sum_inj$INJURIES)
colnames(sum_fi) <- c("cat", "year", "fatalities", "injuries")
head(sum_fi)
```

```
##       cat year fatalities injuries
## 1 tornado 1950         70      659
## 2 tornado 1951         34      524
## 3 tornado 1952        230     1915
## 4 tornado 1953        519     5131
## 5 tornado 1954         36      715
## 6    hail 1955          0        0
```

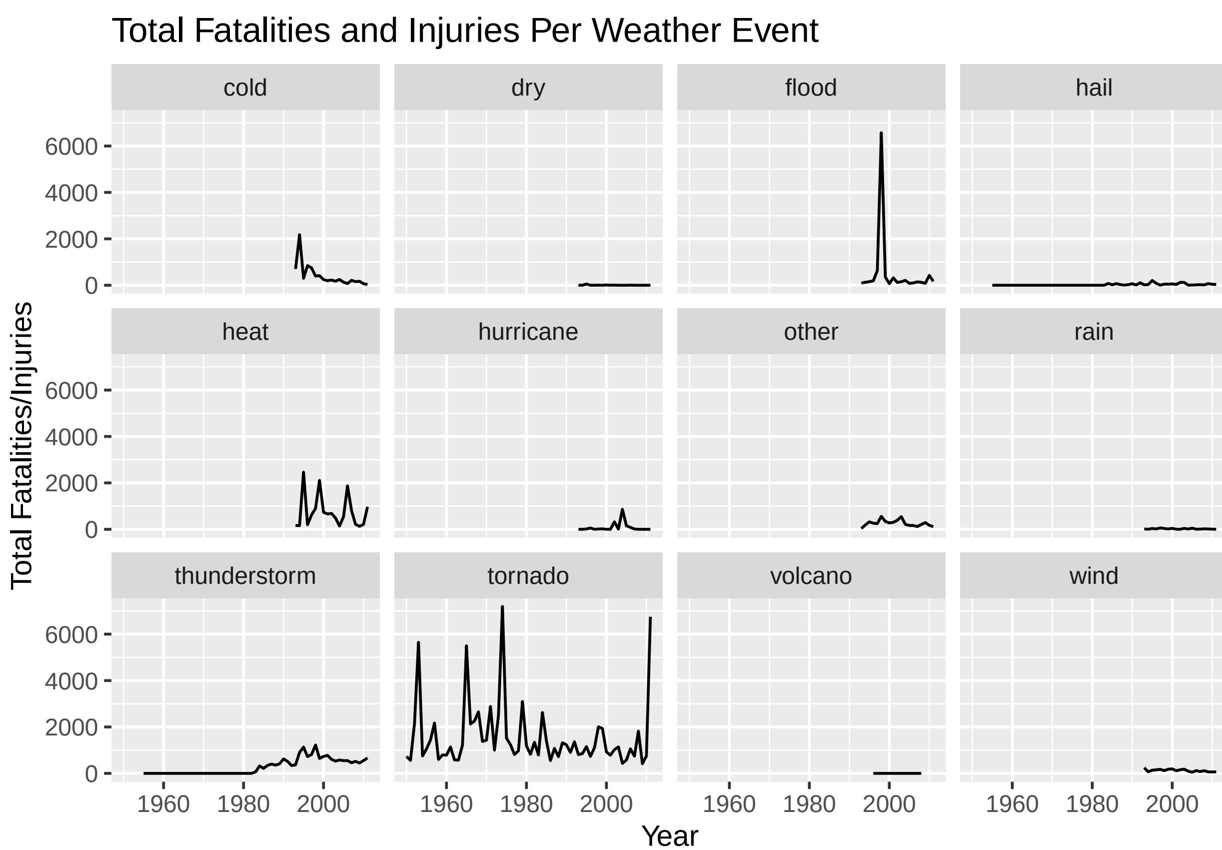
Add a row with the sum of total health impact (fatalities + injuries)

```r
total_fi <- sum_fi
total_fi$total <- total_fi$fatalities + total_fi$injuries
head(total_fi)
```

```
##       cat year fatalities injuries total
## 1 tornado 1950         70      659   729
## 2 tornado 1951         34      524   558
## 3 tornado 1952        230     1915  2145
## 4 tornado 1953        519     5131  5650
## 5 tornado 1954         36      715   751
## 6    hail 1955          0        0     0
```

Then create a plot of total human impact per event each year.

```r
# png(filename= "fi.png", height = 600, width = 600)
ggplot(data = total_fi) +
  geom_line(mapping = aes(x = year, y = total)) +
  facet_wrap(~cat) +
  labs(title = "Total Fatalities and Injuries Per Weather Event",
       x = "Year",
       y = "Total Fatalities/Injuries")
```




```r
# dev.off()
```

You can clearly see that tornados are the most harmful to humans and you can confirm it by ordering the totals:

```
events_fi <- aggregate(total ~ cat, total_fi, sum)
events_fi <- events_fi[order(-events_fi$total), ]
events_fi
```

```
##             cat total
## 10      tornado 97043
## 9  thunderstorm 16346
## 5          heat 14069
## 3         flood 10127
## 1          cold  7504
## 7         other  4905
## 12         wind  2363
## 6     hurricane  1463
## 4          hail  1386
## 8          rain   381
## 2           dry    86
## 11      volcano     0
```

It is also interesting to see when certain categories were starting to be recorded.

Which types of events have the greatest economic consequences?

First, let's take a look at `prop_crop$CROPDMGEXP` and `prop_crop$PROPDMGEXP` to see what's in them and if they need formatting:

```
unique(prop_crop$CROPDMGEXP)
```

```
## [1] ""  "M" "K" "m" "B" "?" "0" "k" "2"
```

```
unique(prop_crop$PROPDMGEXP)
```

```
##  [1] "K" "M" ""  "B" "m" "+" "0" "5" "6" "?" "4" "2" "3" "h" "7" "H" "-" "1" "8"
```

As expected, they need some help.

```
prop_crop2 <- prop_crop
prop_crop2$pexp <- prop_crop2$PROPDMGEXP
prop_crop2$pexp[grep("[Kk]|3", prop_crop2$PROPDMGEXP)] <- 10^3
prop_crop2$pexp[grep("[Mm]|6", prop_crop2$PROPDMGEXP)] <- 10^6
prop_crop2$pexp[grep("[Bb]|9", prop_crop2$PROPDMGEXP)] <- 10^9
prop_crop2$pexp[grep("[Hh]|2", prop_crop2$PROPDMGEXP)] <- 10^2
prop_crop2$pexp[grep("8", prop_crop2$PROPDMGEXP)] <- 10^8
prop_crop2$pexp[grep("5", prop_crop2$PROPDMGEXP)] <- 10^5
prop_crop2$pexp[grep("4", prop_crop2$PROPDMGEXP)] <- 10^4
prop_crop2$pexp[grep("7", prop_crop2$PROPDMGEXP)] <- 10^7
prop_crop2$pexp[grep("0", prop_crop2$PROPDMGEXP)] <- 1
prop_crop2$pexp[grep("1", prop_crop2$PROPDMGEXP)] <- 10
```

```r
prop_crop2$pexp[grep("-", fixed = TRUE, prop_crop2$PROPDMGEXP)] <- 1
prop_crop2$pexp[grep("?", fixed = TRUE, prop_crop2$PROPDMGEXP)] <- 1
prop_crop2$pexp[grep("+", fixed = TRUE, prop_crop2$PROPDMGEXP)] <- 1
prop_crop2$pexp <- as.numeric(prop_crop2$pexp)
```

```r
prop_crop2$cexp <- prop_crop2$CROPDMGEXP
prop_crop2$cexp[grep("[Kk]", prop_crop2$CROPDMGEXP)] <- 10^3
prop_crop2$cexp[grep("[Mm]", prop_crop2$CROPDMGEXP)] <- 10^6
prop_crop2$cexp[grep("[Bb]", prop_crop2$CROPDMGEXP)] <- 10^9
prop_crop2$cexp[grep("0", prop_crop2$CROPDMGEXP)] <- 1
prop_crop2$cexp[grep("?", fixed = TRUE, prop_crop2$CROPDMGEXP)] <- 1
prop_crop2$cexp[grep("2", prop_crop2$CROPDMGEXP)] <- 10^2
unique(prop_crop2$cexp)
```

```
## [1] ""      "1e+06" "1000"  "1e+09" "1"     "100"
```

```r
prop_crop2$cexp <- as.numeric(prop_crop2$cexp)
```

Take out the rows that have `NA`s for both of the created exponential columns

```r
dmg <- prop_crop2[(!(is.na(prop_crop2$pexp) & is.na(prop_crop2$cexp))), ]
```

Next it's time to calculate the totals of crop and property damage by multiplying the `DMG` columns by the `DMGEXP` columns. Then change the `NA`s to 0 so they columns can be added.

```r
pc4 <- dmg
pc4$prop_total <- pc4$PROPDMG * pc4$pexp
pc4$crop_total <- pc4$CROPDMG * pc4$cexp
pc4$crop_total[is.na(pc4$crop_total)] <- 0
pc4$prop_total[is.na(pc4$prop_total)] <- 0
pc4$total <- pc4$prop_total + pc4$crop_total
pc4 <- pc4[ , c(1,2,7,8,13)]
str(pc4)
```

```
## 'data.frame':    440681 obs. of  5 variables:
##  $ STATE : chr  "AL" "AL" "AL" "AL" ...
##  $ EVTYPE: chr  "tornado" "tornado" "tornado" "tornado" ...
##  $ YEAR  : num  1950 1950 1951 1951 1951 ...
##  $ CAT   : chr  "tornado" "tornado" "tornado" "tornado" ...
##  $ total : num  25000 2500 25000 2500 2500 2500 2500 2500 25000 25000 ...
```

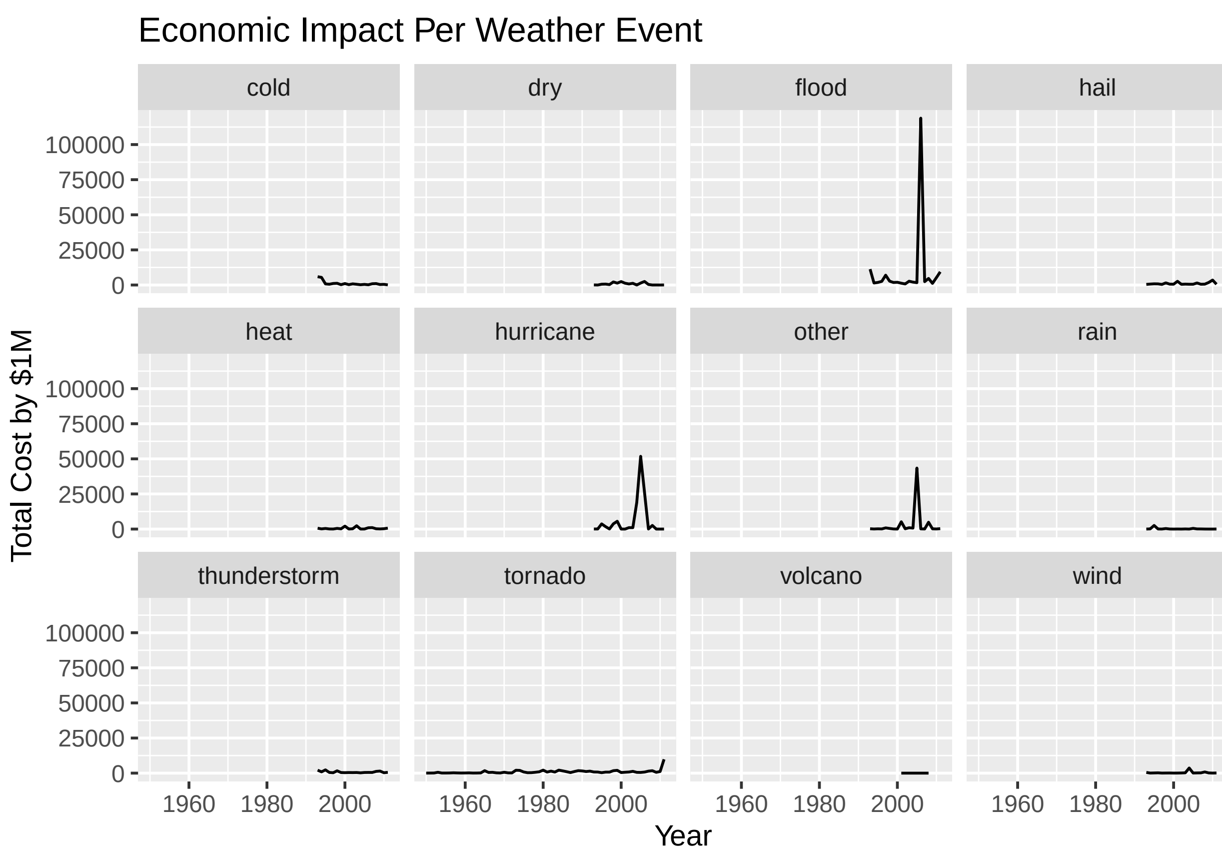
Find the sum per category and divide the total by 1M so it will be more readable in the graph.

```r
event_dmg <- aggregate(total ~ CAT + YEAR, pc4, sum)
event_dmg$total <- event_dmg$total/10^6
str(event_dmg)
```

```
## 'data.frame':    254 obs. of  3 variables:
##  $ CAT  : chr  "tornado" "tornado" "tornado" "tornado" ...
##  $ YEAR : num  1950 1951 1952 1953 1954 ...
##  $ total: num  34.5 65.5 94.1 596.1 85.8 ...
```

Let's graph it

```
# png(filename= "cost.png", height = 600, width = 600)
ggplot(data = event_dmg) +
  geom_line(mapping = aes(x = YEAR, y = total)) +
  facet_wrap(~CAT) +
  labs(title = "Economic Impact Per Weather Event",
       x = "Year",
       y = "Total Cost by $1M")
```




```
# dev.off()
```

```
total_dmg <- aggregate(total ~ CAT, event_dmg, sum)
total_dmg <- total_dmg[order(-total_dmg$total), ]
total_dmg
```

```
##              CAT      total
## 3          flood 180574.433
## 6      hurricane  90271.473
## 7          other  57955.736
## 10       tornado  57418.279
## 1           cold  21341.398
## 4           hail  19024.452
## 2            dry  15025.675
## 9   thunderstorm  15007.062
```

```
## 5          heat   9829.459
## 12         wind   6841.532
## 8          rain   4039.060
## 11      volcano      0.500
```

Here you can see that flooding has been the most costly weather event

Where should you live?

I thought it would be interesting to see which states are the best and worst to live in regarding weather events. Since tornadoes and floods are the most dangerous, I have split the data to contain only those categories.

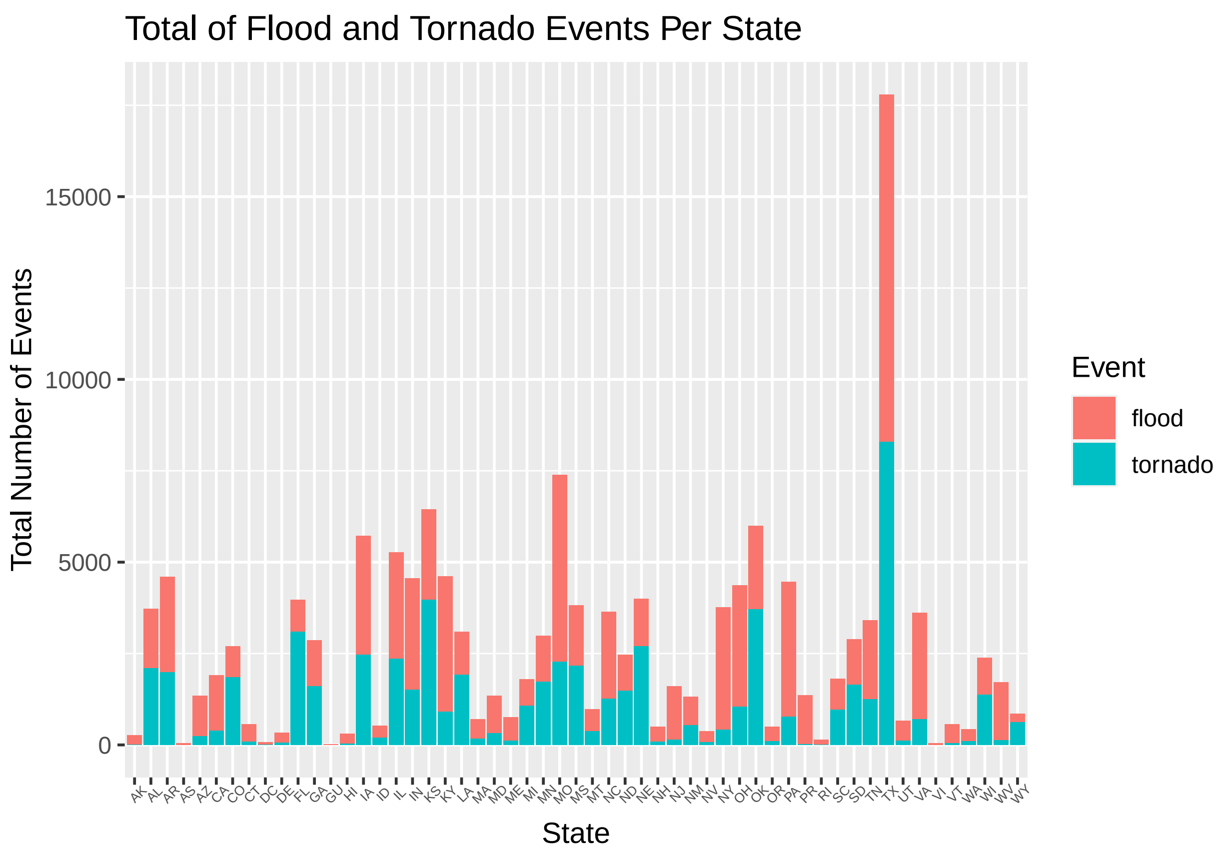
```
states <- prop_crop
states <- states[states$CAT == "tornado" | states$CAT == "flood", c(1, 7, 8)]
head(states)
```

```
##   STATE YEAR     CAT
## 1    AL 1950 tornado
## 2    AL 1950 tornado
## 3    AL 1951 tornado
## 4    AL 1951 tornado
## 5    AL 1951 tornado
## 6    AL 1951 tornado
```

```
states[ , 1] <- as.factor(states$STATE)
states[ , 3] <- as.factor(states$CAT)
summary(states)
```

```
##      STATE            YEAR          CAT
##  TX     :17789   Min.   :1950   flood  :82707
##  MO     : 7391   1st Qu.:1993   tornado:60698
##  KS     : 6443   Median :2001
##  OK     : 6002   Mean   :1996
##  IA     : 5728   3rd Qu.:2007
##  IL     : 5268   Max.   :2011
##  (Other):94784
```

```
ggplot(data = states) +
  geom_histogram(mapping = aes(x = STATE, fill = CAT), stat = "count") +
  labs(title = "Total of Flood and Tornado Events Per State",
       x = "State",
       y = "Total Number of Events",
       fill = "Event") +
  theme(axis.text.x = element_text(angle = 45, size = 4.5))
```



It looks like Texas is the most dangerous state to live in. We can confirm that by ordering the number of events for each event type.

```
flood <- states[states$CAT == "flood", ]
tornado <- states[states$CAT == "tornado", ]
```

```
t <- tornado %>%
  count(STATE)
t <- t[order(t$n), ]
head(t, 3)
```

```
##    STATE  n
## 8     DC  1
## 1     AK  3
## 41    RI 10
```

```
tail(t, 3)
```

```
##    STATE    n
## 37    OK 3709
## 17    KS 3973
## 45    TX 8292
```

```r
f <- flood %>%
  count(STATE)
f <- f[order(f$n), ]
head(f, 3)
```

```
##    STATE  n
## 13    GU 16
## 4     AS 45
## 50    VI 45
```

```r
tail(f, 3)
```

```
##    STATE    n
## 20    KY 3705
## 27    MO 5109
## 47    TX 9497
```

Results

Tornadoes result in the most fatalities and injuries while flooding has the greatest economic impact. Ergo, stay away from areas with tornadoes and floods.
The safest states/districts/territories from tornadoes are:
1. Washington DC
2. Alaska
3. Rhode Island

The most dangerous states/districts/territories for tornadoes are:
1. Texas
2. Kansas
3. Oklahoma

The safest states/districts/territories from floods are:
1. Guam
2. American Samoa
3. Virgin Islands

The most dangerous states/districts/territories for tornadoes are:
1. Texas
2. Missouri
3. Kentucky

We learned that Texas is a very dangerous state to live in and Rhode Island is a nice and safe little state.