Adverse Weather Implication in the USA 1950 - 2011

Abiyu Giday

September 25, 2015

Contents

Abstract: **1**

Data Processing: **1**

Data tydying section: 3

Result section: **4**

Health impact: 4

Economic impact: 5

Other Weather event Factoids from the NOAA dataset: **8**

Overall observation & recomendation: **11**

Abstract:

Every year sever weather events such as storm, flood and tornado adversely impact the United States (US) economy and put the general public's health at greater risk. Analysis contained in this document explores the impact of adverse weather events in the US from 1950 thru 2011. The storm database was obtained from National Oceanic & Atmospheric Administration (NOAA). The database tracks characteristics of major weather events including when/where they occur, financial impact estimation and injury/fatality counts. The database contains 985 different types of weather condition and 902 thousand distinct observations. Therefor understanding and preparing for weather events, to the extent possible, is very important both for the economy and to minimize impact on public health.

The result discussed in this paper attempts to answer the following two questions:-

- Across the United States, which types of weather related events are most harmful with respect to population health?
- Across the United States, which types of weather related events have the greatest economic consequences?

The findings and observation on this paper are intneded to help government and municipal managers who are responsible for planning and prioritizing resources in the events of adverse weather conditions.

Data Processing:

To process the data R static programming language was used in Rstudio Integrated Development Environment(IDE). The R packages that were used to process this data were: ***"dplyr", "ggplot", "tidyr",***

"knitr" and "lubridate". The raw data for this analysis was obtained from NOAA and can be dowloaded frrom here.

The following R script downloads dataset from NOAA and saves it in a *data* direcotry.

```
library(ggplot2)
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
```

```
library(tidyr)
library(lubridate)
library(knitr)
setwd("~/Documents/Data-Science/DataScienceSpecialization/ReproducibleResearch/proj2/Reproduceable")

if(!file.exists("./data")){dir.create("./data")}
fileUrl <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
download.file(fileUrl, destfile = "./data/StormData.csv.bz2", method = "curl")

strmd <- tbl_df(read.csv(bzfile("./data/StormData.csv.bz2")))
str(strmd)
```

```
## Classes 'tbl_df', 'tbl' and '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 ...
```

Data tydying section:

The following code clean the data and set it in a format ready for the analysis

```
sd1 <- select(strmd, BGN_DATE, STATE, EVTYPE, COUNTY, COUNTYNAME, F, FATALITIES, INJURIES,   PROPDMG, C
sd1$BGN_DATE <- as.Date(sd1$BGN_DATE, format = "%m/%d/%Y %H:%M:%S")
sd1 <- separate(sd1, BGN_DATE, c("year", "month", "day"), sep = "-")

sd1
```

```
## Source: local data frame [902,297 x 14]
##
##     year month   day  STATE  EVTYPE COUNTY COUNTYNAME     F FATALITIES
##    (chr) (chr) (chr) (fctr)  (fctr)  (dbl)     (fctr) (int)      (dbl)
## 1   1950    04    18     AL TORNADO     97     MOBILE     3          0
## 2   1950    04    18     AL TORNADO      3    BALDWIN     2          0
## 3   1951    02    20     AL TORNADO     57    FAYETTE     2          0
## 4   1951    06    08     AL TORNADO     89    MADISON     2          0
## 5   1951    11    15     AL TORNADO     43    CULLMAN     2          0
## 6   1951    11    15     AL TORNADO     77 LAUDERDALE     2          0
## 7   1951    11    16     AL TORNADO      9     BLOUNT     2          0
## 8   1952    01    22     AL TORNADO    123 TALLAPOOSA     1          0
## 9   1952    02    13     AL TORNADO    125 TUSCALOOSA     3          1
## 10  1952    02    13     AL TORNADO     57    FAYETTE     3          0
## ..   ...   ...   ...    ...     ...    ...        ...   ...        ...
## Variables not shown: INJURIES (dbl), PROPDMG (dbl), CROPDMG (dbl),
##   LATITUDE (dbl), LONGITUDE (dbl)
```

Explanation on data filter in the codes it is important to point out the reason behind setting the filters for this analysis, and why high and low water mark cut off numbers were selected for both health and economic dataset. The raw data from NOAA contains overwhelmingly more Tornado observations than any other weather event from 1950 thru 1990. Tornado's alone account for 80% of the total adverse weather counts in the NOAA dataset, which distorts the finding without the filters.

Result section:

Health impact:

In terms of health impact on the population, the analysis shows that, ***Tornado***, accounts for 80.2% of the total injuries and fatalities for years spanning from 1950 to 2011. Tornado comes 4th, in frequency of occurrence, after "Hail", "TSTM wind" and "Thunder storm", but it was the cause of 55,464 injuries and fatalities. The second weather event that impacted US population's health the most was **Excessive heat** causing harm to 4,265 individuals (6.2%), and the 3rd most cause of injury and fatality was **Flood** harming 2, 562 people (3.7%).

Here are the codes used to create a data frame for the health impact analysis.

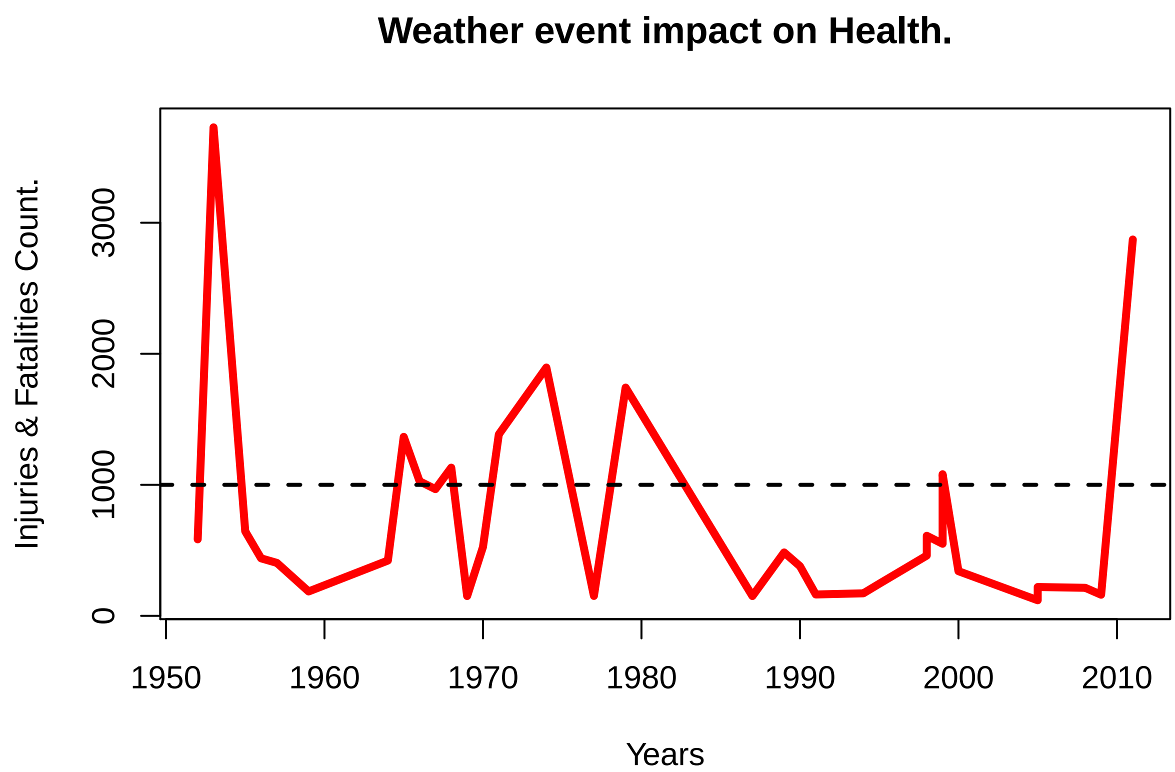
```
# This code adds a column that combines the Injuries and Fatality for each weather event type.
df1 <- sd1  %>% filter(INJURIES > 100 & FATALITIES > 10 )  %>%
                 group_by(year, EVTYPE)  %>%
                       summarise_each(funs(sum), INJURIES, FATALITIES)  %>%
                                mutate( TotalHealthImpact = INJURIES + FATALITIES) %>%
                                   arrange(TotalHealthImpact)
df1
```

```
## Source: local data frame [33 x 5]
## Groups: year [30]
##
##     year  EVTYPE INJURIES FATALITIES TotalHealthImpact
##    (chr)  (fctr)    (dbl)      (dbl)             (dbl)
## 1   1952 TORNADO      505         79               584
## 2   1953 TORNADO     3339        389              3728
## 3   1955 TORNADO      550         95               645
## 4   1956 TORNADO      400         39               439
## 5   1957 TORNADO      356         48               404
## 6   1959 TORNADO      175         11               186
## 7   1964 TORNADO      389         33               422
## 8   1965 TORNADO     1271         95              1366
## 9   1966 TORNADO      954         73              1027
## 10  1967 TORNADO      910         57               967
## ..   ...     ...      ...        ...               ...
```

Here is the graph for Most Health impact in the US

This graph shows the combined weather evetns that caused at list 100 injuries and 10 fatalities.

```
plot(df1$year, df1$TotalHealthImpact, type = "l", lwd = 4, col = "red", main = "Weather event impact on
abline(h = 1000, lwd = 2, lty =2)
```



Here is the ranking of weather event type per total number of injuries and fatalities. *Note: the ranking numbers are based on the filter window set in the script*

```
df11 <- df1 %>%
             group_by(EVTYPE)  %>%
                summarise_each(funs(sum), TotalHealthImpact)  %>%
                       arrange(desc(TotalHealthImpact))

df11
```

```
## Source: local data frame [6 x 2]
##
##                EVTYPE TotalHealthImpact
##                (fctr)             (dbl)
## 1             TORNADO             22756
## 2      EXCESSIVE HEAT              1081
## 3               FLOOD               611
## 4                HEAT               245
## 5             TSUNAMI               161
## 6 HURRICANE/TYPHOON                 119
```

Economic impact:

The weather event with the most adverse economic consequences was *Tornado*, costing US economy $3.3 billion dollars. The second weather condition that had dire economic impact was *Flash* Flood" at a cost of $1.6 billion dollars. *TSM WIND* was the 4th most costly weather event at $1.4 billion dollars.

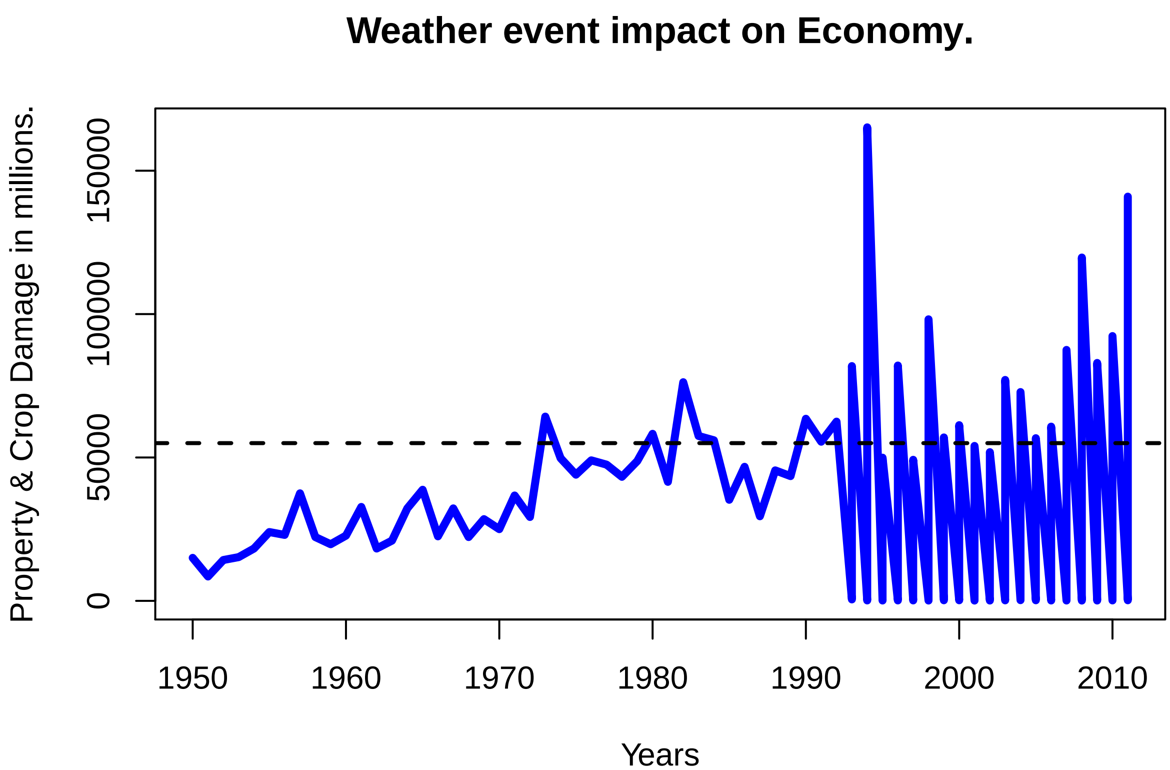
Here are the code used to to create a data frame for the economic impact analysis.

```r
#This code adds a column that combines the property and crop damage costs.
df18 <- sd1  %>% filter(PROPDMG > 100 | CROPDMG > 100)  %>%
                 group_by(year, EVTYPE)  %>%
                       summarise_each(funs(sum), PROPDMG, CROPDMG)  %>%
                                mutate( TotalEconthImpact = PROPDMG + CROPDMG) %>%
                                                             arrange(TotalEconthImpact)
#df18
```

Here is the graph for Most Economic impact in the US.

The filter for this graph is set from a minimum of $100 thousand dollars worth of property or crop damage.

```r
plot(df18$year, df18$TotalEconthImpact, type = "l", lwd = 4, col = "blue", main = "Weather event impact
abline(h = 55000, lwd = 2, lty =2)
```



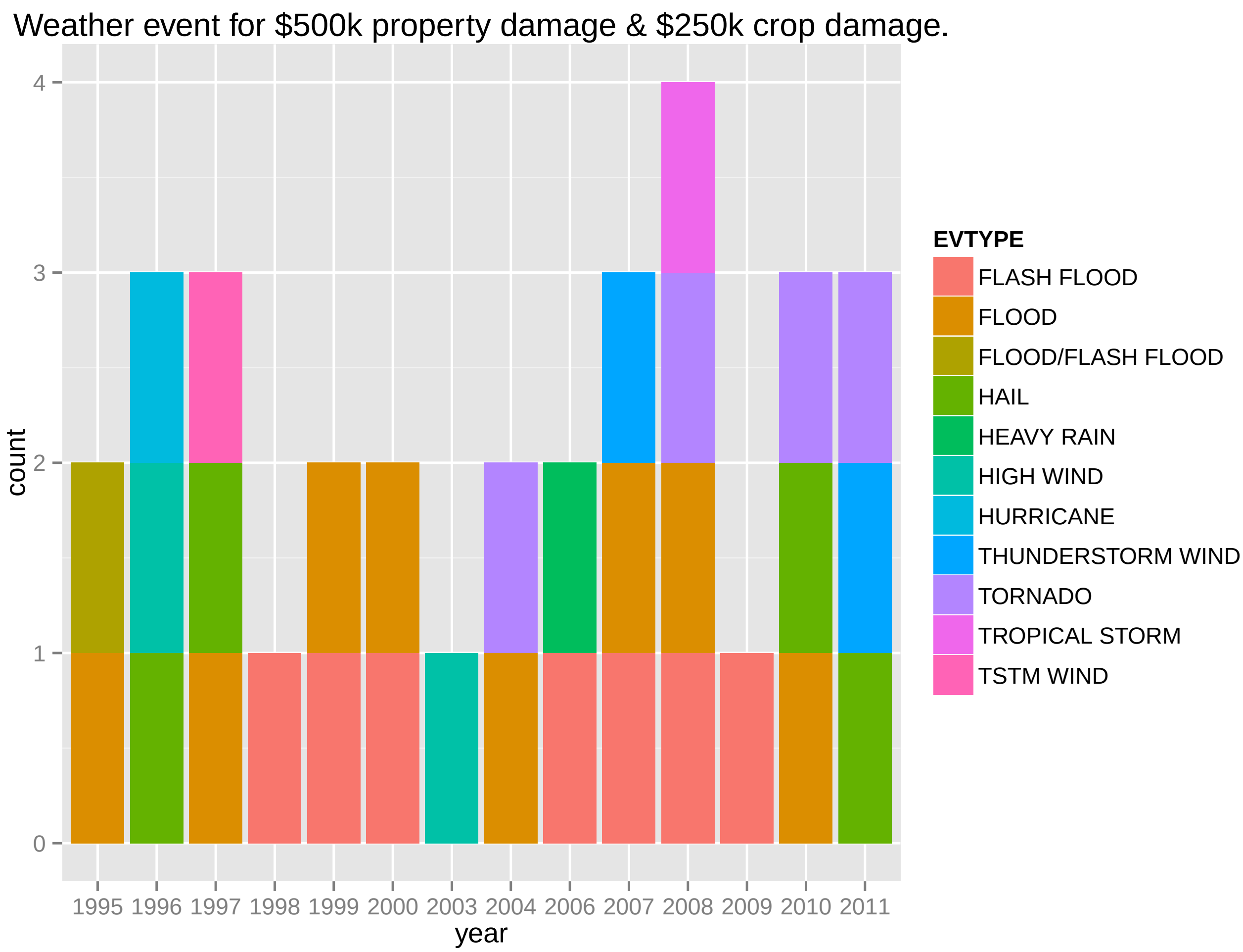
Here is another graph that shows weather events with Most Economic impact from 1995 to 2011

The following graphs shows the impact on the economy for property damages more than $500k & crop damage more than $250k. Because more diverse weather events were collected in the 1990's, the graph reflects more variability from the 1990's to 2011.

```r
# The following filter is setup to examine Economic impact from 1995
df19 <- sd1  %>% filter(PROPDMG > 500 & CROPDMG > 250)  %>%
                 group_by(year, EVTYPE)  %>%
                       summarise_each(funs(sum), PROPDMG, CROPDMG)  %>%
                                mutate( TotalEconthImpact = PROPDMG + CROPDMG) %>%
```

```
                                              arrange(TotalEconthImpact)
#df19


k <- ggplot(df19, aes(year, fill=EVTYPE)) # bar...
k + geom_bar() + ggtitle("Weather event for $500k property damage & $250k crop damage.")
```



Here is the ranking of economic impact per weather event for the combined property and crop damages. *Note: the ranking numbers are based on the filter window set during the data frame creation.*

```
#Filter weather evetns that cost the economy more than 0 for both perpoerty and crop damage.
df1 <- sd1  %>% filter(PROPDMG > 0 | CROPDMG > 0)  %>%
                 group_by(year, EVTYPE)  %>%
                       summarise_each(funs(sum), PROPDMG, CROPDMG)  %>%
                                mutate( TotalEconthImpact = PROPDMG + CROPDMG)

# rank weather event with total financial impact.
df11 <- df1 %>%
             group_by(EVTYPE)  %>%
                summarise_each(funs(sum), TotalEconthImpact)  %>%
                       arrange(desc(TotalEconthImpact))

df11
```

```
## Source: local data frame [431 x 2]
##
##                EVTYPE TotalEconthImpact
##                (fctr)             (dbl)
## 1             TORNADO         3312276.7
## 2         FLASH FLOOD         1599325.1
## 3           TSTM WIND         1445168.2
## 4                HAIL         1268289.7
## 5               FLOOD         1067976.4
## 6   THUNDERSTORM WIND          943635.6
## 7           LIGHTNING          606932.4
## 8  THUNDERSTORM WINDS          464978.1
## 9           HIGH WIND          342014.8
## 10       WINTER STORM          134699.6
## ..                ...               ...
```

Other Weather event Factoids from the NOAA dataset:

The maximum number of Fatalties, Injuries, Property damage and crop damage

```r
max(sd1$FATALITIES)
```

```
## [1] 583
```

```r
max(sd1$INJURIES)
```

```
## [1] 1700
```

```r
max(sd1$CROPDMG)
```

```
## [1] 990
```

```r
max(sd1$PROPDMG)
```

```
## [1] 5000
```

```r
sd1  %>% filter( FATALITIES == "583" | INJURIES == "1700"| CROPDMG == "990" | PROPDMG == "5000")
```

```
## Source: local data frame [7 x 14]
##
##    year month   day  STATE            EVTYPE COUNTY
##   (chr) (chr) (chr) (fctr)            (fctr)  (dbl)
## 1  1979    04    10     TX           TORNADO    485
## 2  1995    07    12     IL              HEAT      0
## 3  2004    05    01     MT           DROUGHT     24
## 4  2009    07    26     NC THUNDERSTORM WIND     69
## 5  2010    05    13     IL       FLASH FLOOD    131
## 6  2010    05    13     IL       FLASH FLOOD     73
## 7  2011    10    29     AM        WATERSPOUT    555
## Variables not shown: COUNTYNAME (fctr), F (int), FATALITIES (dbl),
##   INJURIES (dbl), PROPDMG (dbl), CROPDMG (dbl), LATITUDE (dbl), LONGITUDE
##   (dbl)
```

Most and list frequent weather events.

```
df11 <- sd1 %>% group_by(EVTYPE)  %>%  summarise(count = n()) %>% arrange(desc(count))

head(df11)
```

```
## Source: local data frame [6 x 2]
##
##              EVTYPE  count
##              (fctr)  (int)
## 1              HAIL 288661
## 2         TSTM WIND 219940
## 3 THUNDERSTORM WIND  82563
## 4           TORNADO  60652
## 5       FLASH FLOOD  54277
## 6             FLOOD  25326
```

```
tail(df11)
```

```
## Source: local data frame [6 x 2]
##
##                    EVTYPE count
##                    (fctr) (int)
## 1               WIND/HAIL     1
## 2 WINTER STORM HIGH WINDS     1
## 3  WINTER STORM/HIGH WIND     1
## 4 WINTER STORM/HIGH WINDS     1
## 5             Wintry Mix     1
## 6                     WND     1
```

The years with most and list active weather events listed.

```
df12 <- sd1 %>% group_by(year)  %>%  summarise(count = n()) %>% arrange(desc(count))
head(df12)
```

```
## Source: local data frame [6 x 2]
##
##    year count
##   (chr) (int)
## 1  2011 62174
## 2  2008 55663
## 3  2010 48161
## 4  2009 45817
## 5  2006 44034
## 6  2007 43289
```

```
tail(df12)
```

```
## Source: local data frame [6 x 2]
##
##    year count
##   (chr) (int)
## 1  1955  1413
## 2  1954   609
## 3  1953   492
## 4  1952   272
## 5  1951   269
## 6  1950   223
```

States that experienced the most and list weather events

```r
df13 <- sd1 %>% group_by(STATE)  %>%  summarise(count = n()) %>% arrange(desc(count))
head(df13)
```

```
## Source: local data frame [6 x 2]
##
##    STATE count
##   (fctr) (int)
## 1     TX 83728
## 2     KS 53440
## 3     OK 46802
## 4     MO 35648
## 5     IA 31069
## 6     NE 30271
```

```r
tail(df13)
```

```
## Source: local data frame [6 x 2]
##
##    STATE count
##   (fctr) (int)
## 1     PK    23
## 2     SL     7
## 3     XX     2
## 4     MH     1
## 5     PM     1
## 6     ST     1
```

Months with the most and list Weather events

```r
df14 <- sd1 %>% group_by(month)  %>%  summarise(count = n()) %>% arrange(desc(count))
head(df14)
```

```
## Source: local data frame [6 x 2]
##
##   month  count
##   (chr)  (int)
```

```
## 1    06 174450
## 2    05 150159
## 3    07 136811
## 4    04 100371
## 5    08  96424
## 6    03  55246
```

```
tail(df14)
```

```
## Source: local data frame [6 x 2]
##
##   month count
##   (chr) (int)
## 1    09 44374
## 2    02 32608
## 3    01 31025
## 4    10 28464
## 5    11 26545
## 6    12 25820
```

Total Fatality, Injury count & Total Property and Crop cost incured.

```
 sum(sd1$FATALITIES)
```

```
## [1] 15145
```

```
 sum(sd1$INJURIES)
```

```
## [1] 140528
```

```
 sum(sd1$CROPDMG) * 1000
```

```
## [1] 1377827320
```

```
 sum(sd1$PROPDMG) * 1000
```

```
## [1] 10884500010
```

Overall observation & recomendation:

The US sever weather data analysis for the span of 61 years shows that 97.6% of the the weather events didn't cause health problems or had meaningfully measurable economic impact. However, the data analysis on this paper shows, 2.4% of the sever weather events had tremendous cost to public health, and had a significant negative consequences to the US economy.

Here is a list that puts the total counts and costs in numbers:

15,145 fatalities were incurred.

140, 528 injuries were caused.

$1.4 billion was the price tag for crop damages.

$10.9 billions in property damages.

The weather event that caused the maximum number of injuries took place on the April 10, 1979 ***Tornado*** in Wichita county Texas. There were 1700 reported injures. ***Heat*** caused the most fatality at 583. This weather event happened on July 12, 1995 in Illinois. In terms of economic impact, there were four weather events that caused the most property damage at the price tag of 5 million each. Two of the four were caused by ***Flash Flood*** in Illinois, Mercer and Henry county on May 13, 2010. North Carolina's Franklin county was the third county with property damage of 5 million dollars caused by ***Thunderstorm*** on July 26, 2009. A ***Waterspout*** event on marine zone 555, located in Melbroune, Florida, that took place on October 29, 2011 also had a 5 million dollar property damage. The sustained ***Draught*** weather condition caused 990 thousand dollars in the state of Montana recorded on May 1, 2004.

While there isn't full proof way to stop mother nature, patterns observed from this analysis can be used to prepare and align rescues to help minimize the damage. For example, ***April-May-June-July*** are the months when sever weather events tend to occur the most. Thus, planning events or planting crops in the states should factor the pattern and plan accordingly.

Note: The database from NOAA contains variables that were not explored on this analysis, adding those variables to the analysis could result in additional insight that will help municipalities save lives and plan their budget.