Exploring Severe Weather Events' Impacts on Public Health and Economy across the United States

Xinyu W

Synopsis

In this report we aim to investigate the impacts of different types of severe weather events across the United States with a focus on public health and national economy.

We obtained data from the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database, which tracks characteristics of major storms and weather events in the United States from year 1950 to November 2011. This includes when and where the events occurred, as well as estimates of any fatalities, injuries and property damage.

From these data, we found that, **tornado, heat and wind** are most harmful with respect to population health, while **flood, hurricane and storm** have the greatest economic consequences.

Preparation Works

Set the global options

```
library(knitr)
opts_chunk$set(fig.path = "Figs/", warning=FALSE, message = FALSE, echo=TRUE)
options(scipen = 999)
```

Loading the Raw Data

From the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database, we obtained information of major storms and weather events across the U.S. We obtained the files from year 1950 to November 2011.

There is also some documentation of the database available. Here you will find how some of the variables are constructed/defined.

- National Weather Service Storm Data Documentation
- National Climatic Data Center Storm Events FAQ

Download, unzip and read in the data

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

```r
if (!exists("NOAA")){
        NOAA <- read.csv(bzfile("./data/NOAA.csv"),header = TRUE)
}
```

Look into some details of the data set

```r
dim(NOAA)
```

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

```r
head(NOAA,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
```

We can see that the data set contains 902297 observations of 37 variables. The events in the database start in the year 1950 and end in November 2011.

In the earlier years of the database there are generally fewer events recorded, most likely due to a lack of good records. More recent years should be considered more complete.

Data Processing

Transformation on the Raw Data

Here, we reinforce the two questions we want to address:

1. Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health?
2. Across the United States, which types of events have the greatest economic consequences?

For the data set, we choose to select the **fatalities** and **injuries** data to analyze the influence on public health, and for national economy, we would concentrate on the estimation of **property and crop damages**. Based on that, we subset the following data out of the original:

```r
library(dplyr)
interestData <- NOAA %>% select(EVTYPE, FATALITIES, INJURIES, PROPDMG, PROPDMGEXP, CROPDMG, CROPDMGEXP)
head(interestData, 6)
```

```
##    EVTYPE FATALITIES INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP
## 1 TORNADO          0       15    25.0          K       0
## 2 TORNADO          0        0     2.5          K       0
## 3 TORNADO          0        2    25.0          K       0
## 4 TORNADO          0        2     2.5          K       0
## 5 TORNADO          0        2     2.5          K       0
## 6 TORNADO          0        6     2.5          K       0
```

NOTES on the variables:

- Common variables
  - EVTYPE: the type of the weather event
- Public health related variables
  - FATALITIES: approximate number of deaths
  - INJURIES: approximate number of injuries
- Economy related variables
  - PROPDMG: estimated property damages
  - PROPDMGEXP: the units of estimated property damages
  - CROPDMG: estimated crop damages
  - CROPDMGEXP: the units of estimated crop damages

Arrange the Event Types

When we look into the event types, we can immediately find some duplicates or some types that can be generated into one type. For example:

```r
sort(unique(interestData$EVTYPE))[1:10]
```

```
##  [1] "   HIGH SURF ADVISORY" " COASTAL FLOOD"        " FLASH FLOOD"
##  [4] " LIGHTNING"            " TSTM WIND"            " TSTM WIND (G45)"
##  [7] " WATERSPOUT"           " WIND"                 "?"
## [10] "ABNORMAL WARMTH"
```

There are several index all containing "Wind" or "Flood." Therefore, we can group some index together.

```r
#insert a new column for grouping
interestData$New.Event <- interestData$EVTYPE
#reorganize the event types
interestData$New.Event[grep("DROUGHT", interestData$EVTYPE, ignore.case = TRUE)] <- "DROUGHT"
interestData$New.Event[grep("COLD", interestData$EVTYPE, ignore.case = TRUE)] <- "COLD"
interestData$New.Event[grep("HAIL", interestData$EVTYPE, ignore.case = TRUE)] <- "HAIL"
interestData$New.Event[grep("HURRICANE", interestData$EVTYPE, ignore.case = TRUE)] <- "HURRICANE"
interestData$New.Event[grep("HEAT", interestData$EVTYPE, ignore.case = TRUE)] <- "HEAT"
interestData$New.Event[grep("FLOOD", interestData$EVTYPE, ignore.case = TRUE)] <- "FLOOD"
```

```
interestData$New.Event[grep("LIGHTNING", interestData$EVTYPE, ignore.case = TRUE)] <- "LIGHTNING"
interestData$New.Event[grep("RAIN", interestData$EVTYPE, ignore.case = TRUE)] <- "RAIN"
interestData$New.Event[grep("TORNADO", interestData$EVTYPE, ignore.case = TRUE)] <- "TORNADO"
interestData$New.Event[grep("WIND", interestData$EVTYPE, ignore.case = TRUE)] <- "WIND"
interestData$New.Event[grep("WINTER", interestData$EVTYPE, ignore.case = TRUE)] <- "WINTER"
interestData$New.Event[grep("WILDFIRE", interestData$EVTYPE, ignore.case = TRUE)] <- "WILDFIRE"
interestData$New.Event[grep("STORM", interestData$EVTYPE, ignore.case = TRUE)] <- "STORM"
interestData$New.Event[grep("SNOW", interestData$EVTYPE, ignore.case = TRUE)] <- "SNOW"

#To show the new data set
sort(table(interestData$New.Event), decreasing = TRUE)[1:10]
```

```
##
##         HAIL         WIND        STORM        FLOOD      TORNADO         SNOW
##       289270       255385       124599        82686        60699        17705
##    LIGHTNING         RAIN       WINTER FUNNEL CLOUD
##        15760        12175         8160         6839
```

Then we can see much clearer of different event types.

Check on Missing Values

Before we do any further processing and analyzing of the data, we first check for any missing values there.

```
sum(is.na(interestData$FATALITIES))
```

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

```
sum(is.na(interestData$INJURIES))
```

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

```
sum(is.na(interestData$PROPDMG))
```

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

```
sum(is.na(interestData$PROPDMGEXP))
```

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

```
sum(is.na(interestData$CROPDMG))
```

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

```
sum(is.na(interestData$CROPDMGEXP))
```

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

From above, we see there are no missing values in our data set. We can then begin our processing.

Impact on Public Health

To evaluate the public health impact:

1. We choose related data (the fatalities and injuries) from interestData.
2. We then summarize the data frame by weather events types.
3. We order the resulting data frame by the sum of fatalities/injuries.

```
fatalData <- interestData %>% select(New.Event, FATALITIES) %>% group_by(New.Event) %>% summarise(sum.fa
head(fatalData, 8)
```

```
## # A tibble: 8 x 2
##   New.Event   sum.fatalities
##   <chr>                <dbl>
## 1 TORNADO               5636
## 2 HEAT                  3138
## 3 FLOOD                 1524
## 4 WIND                  1235
## 5 LIGHTNING              817
## 6 STORM                  633
## 7 RIP CURRENT            368
## 8 AVALANCHE              224
```

```
injuriesData <- interestData %>% select(New.Event, INJURIES) %>% group_by(New.Event) %>% summarise(sum.
head(injuriesData, 8)
```

```
## # A tibble: 8 x 2
##   New.Event sum.injuries
##   <chr>            <dbl>
## 1 TORNADO          91407
## 2 HEAT              9224
## 3 WIND              9001
## 4 FLOOD             8602
## 5 STORM             6692
## 6 LIGHTNING         5231
## 7 HAIL              1371
## 8 HURRICANE         1328
```

Impact on National Economy

1. Interpret special index in the data set

To analyze the economic impacts brought by weather events, we first need to understand the property and crop damages data. Here, we see special index we need to figure out.

```
paste("The characters in PROPDMGEXP include")
```

```
## [1] "The characters in PROPDMGEXP include"
```

```
sort(table(interestData$PROPDMGEXP), decreasing = TRUE)
```

```
##
##             K      M      0      B      5      1      2      ?      m      H
## 465934 424665  11330    216     40     28     25     13      8      7      6
##      +      7      3      4      6      -      8      h
##      5      5      4      4      4      1      1      1
```

```
paste("The characters in CROPDMGEXP include")
```

```
## [1] "The characters in CROPDMGEXP include"
```

```
sort(table(interestData$CROPDMGEXP), decreasing = TRUE)
```

```
##
##             K      M      k      0      B      ?      2      m
## 618413 281832   1994     21     19      9      7      1      1
```

In the National Weather Service Storm Data Documentation, there's one sentence explaining the index in the PROPDMGEXP and CROPDMGEXP:

Alphabetical characters used to signify magnitude include "K" for thousands, "M" for millions, and "B" for billions.

Combined with the trial online, we find the index can be interpreted as follows:

- K or k: thousand (10^3)
- M or m: million (10^6)
- B or b: billion (10^9)
- H or h: hundred (10^2)
- 0,1,2,3,4,5,6,7,8 : 10
- "+" : 1
- "-" : 0
- "?" : 0
- blank: 0

To have a data frame with interpretations of these units:

```
economyData <- interestData %>% select(New.Event, PROPDMG, PROPDMGEXP, CROPDMG, CROPDMGEXP)
#transform the PROPDMGEXP data
economyData$PROPDMGEXP[grep("[Kk]", economyData$PROPDMGEXP, ignore.case = TRUE)] <- 10^3
economyData$PROPDMGEXP[grep("[Mm]", economyData$PROPDMGEXP, ignore.case = TRUE)] <- 10^6
economyData$PROPDMGEXP[grep("[Bb]", economyData$PROPDMGEXP, ignore.case = TRUE)] <- 10^9
economyData$PROPDMGEXP[grep("[Hh]", economyData$PROPDMGEXP, ignore.case = TRUE)] <- 10^2
economyData$PROPDMGEXP[(economyData$PROPDMGEXP == "")] <- 0
economyData$PROPDMGEXP[(economyData$PROPDMGEXP == "?")] <- 0
economyData$PROPDMGEXP[(economyData$PROPDMGEXP == "-")] <- 0
economyData$PROPDMGEXP[(economyData$PROPDMGEXP == "+")] <- 1
economyData$PROPDMGEXP[(economyData$PROPDMGEXP == "0")] <- 10
economyData$PROPDMGEXP[(economyData$PROPDMGEXP == "1")] <- 10
economyData$PROPDMGEXP[(economyData$PROPDMGEXP == "2")] <- 10
```

```
economyData$PROPDMGEXP[(economyData$PROPDMGEXP == "3")] <- 10
economyData$PROPDMGEXP[(economyData$PROPDMGEXP == "4")] <- 10
economyData$PROPDMGEXP[(economyData$PROPDMGEXP == "5")] <- 10
economyData$PROPDMGEXP[(economyData$PROPDMGEXP == "6")] <- 10
economyData$PROPDMGEXP[(economyData$PROPDMGEXP == "7")] <- 10
economyData$PROPDMGEXP[(economyData$PROPDMGEXP == "8")] <- 10

#transform the CROPDMGEXP data
economyData$CROPDMGEXP[grep("[Kk]", economyData$CROPDMGEXP, ignore.case = TRUE)] <- 10^3
economyData$CROPDMGEXP[grep("[Mm]", economyData$CROPDMGEXP, ignore.case = TRUE)] <- 10^6
economyData$CROPDMGEXP[grep("[Bb]", economyData$CROPDMGEXP, ignore.case = TRUE)] <- 10^9
economyData$CROPDMGEXP[grep("[Hh]", economyData$CROPDMGEXP, ignore.case = TRUE)] <- 10^2
economyData$CROPDMGEXP[(economyData$CROPDMGEXP == "?")] <- 0
economyData$CROPDMGEXP[(economyData$CROPDMGEXP == "0")] <- 10
economyData$CROPDMGEXP[(economyData$CROPDMGEXP == "2")] <- 10
economyData$CROPDMGEXP[(economyData$CROPDMGEXP == "")] <- 0

#To show the new unit variables
paste("The units in PROPDMGEXP include")
```

```
## [1] "The units in PROPDMGEXP include"
```

```
sort(table(economyData$PROPDMGEXP), decreasing = TRUE)
```

```
##
##         10       1000    1000000 1000000000        100
##     466248     424665      11337          40          7
```

```
paste("The units in CROPDMGEXP include")
```

```
## [1] "The units in CROPDMGEXP include"
```

```
sort(table(economyData$CROPDMGEXP), decreasing = TRUE)
```

```
##
##          0       1000    1000000         10 1000000000
##     618413     281853       1995         27          9
```

2. Group by the event types and re-arrange the damages data

- Generate the property damages data frame by weather event types

```
economyData <- mutate(economyData, propCost = PROPDMG * as.numeric(PROPDMGEXP))
propDMG <- economyData %>% select(New.Event, propCost) %>% group_by(New.Event) %>% summarise(sum.PropCos
#To show the property damages data
head(propDMG,8)
```

```
## # A tibble: 8 x 2
##   New.Event  sum.PropCost
```

```
##   <chr>             <dbl>
## 1 FLOOD     167502199413
## 2 HURRICANE  84656180010
## 3 STORM      73054022622
## 4 TORNADO    56993100717
## 5 HAIL       15733046447
## 6 WIND       12454677314.
## 7 WILDFIRE    4865614000
## 8 RAIN        3254491210
```

- Generate the crop damages data frame by weather events types

```
economyData <- mutate(economyData, cropCost = CROPDMG * as.numeric(CROPDMGEXP))
cropDMG<- economyData %>% select(New.Event, cropCost) %>% group_by(New.Event) %>% summarise(sum.CropCos
#To show the crop damages data
head(cropDMG,8)
```

```
## # A tibble: 8 x 2
##   New.Event    sum.CropCost
##   <chr>               <dbl>
## 1 DROUGHT       13972566000
## 2 FLOOD         12266906100
## 3 STORM          6406919600
## 4 HURRICANE      5505292800
## 5 HAIL           3046837650
## 6 WIND           1519029150
## 7 COLD           1409115500
## 8 FROST/FREEZE   1094086000
```

Now we have our processed data ready to make plots, analyze and achieve the results.

Results

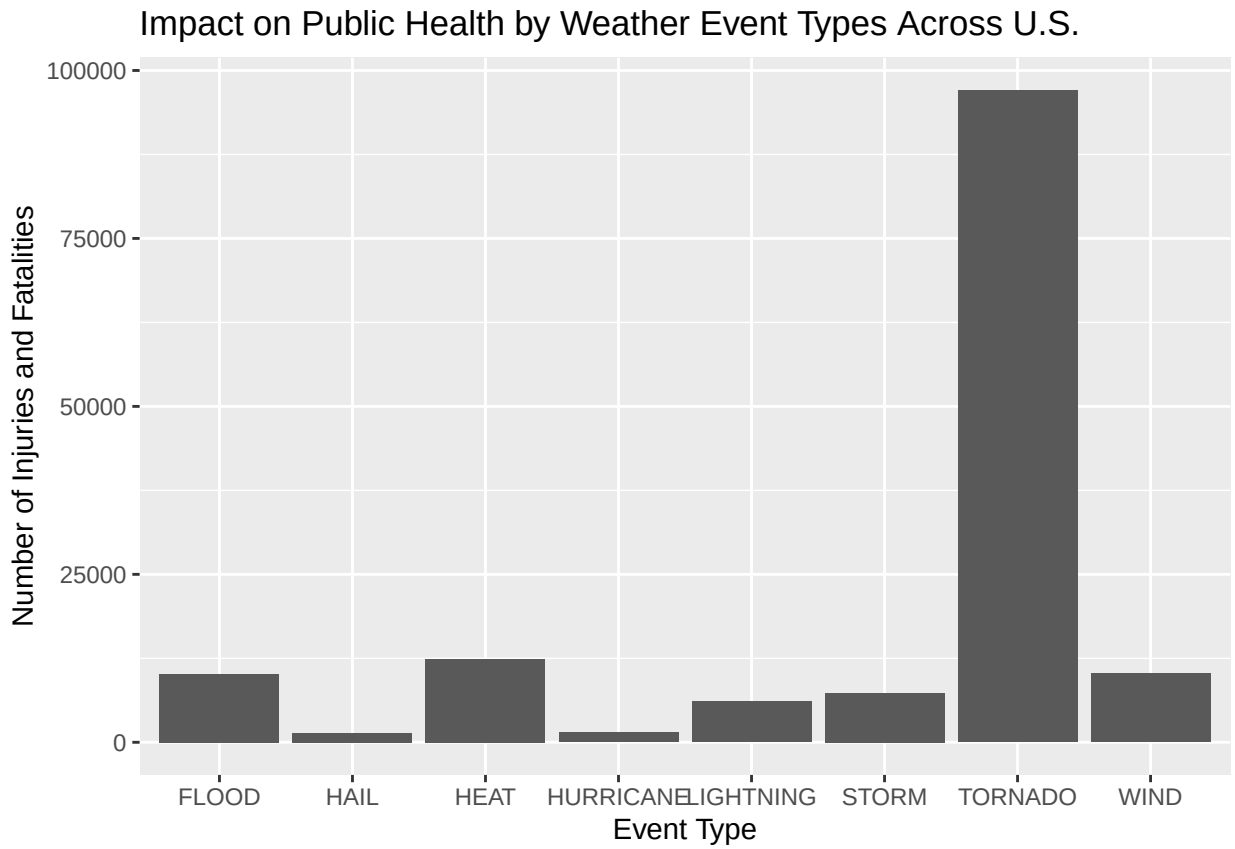
Q1: Across the United States, which types of events are most harmful with respect to population health?

In order to show impacts on public health by different event types, we can make a barplot of sum number of fatalities and injuries from year 1950 to November 2011.

- Merge the data sets related to public health by event types
- Conduct the plot using ggplot2 system
- Generate the result

```
#re-format the health related data
healthResult <- inner_join(fatalData, injuriesData, by = "New.Event")
healthResult <- healthResult %>% mutate(healthResult, Total=sum.fatalities+sum.injuries) %>% arrange(-T

library(ggplot2)
#create the plot
ggplot(healthResult[1:8,], aes(x=New.Event, y=Total))+geom_bar(stat = "identity")+xlab("Event Type")+yl
```




```
head(healthResult,3)
```

```
## # A tibble: 3 x 4
##   New.Event sum.fatalities sum.injuries Total
##   <chr>              <dbl>        <dbl> <dbl>
## 1 TORNADO             5636        91407 97043
## 2 HEAT                3138         9224 12362
## 3 WIND                1235         9001 10236
```

The plot combined with the table shows us that the top three weather events with greatest impacts on public health are **tornado, heat and wind.** And tornado has the obvious greatest influence.

Q2: Across the United States, which types of events have the greatest economic consequences?

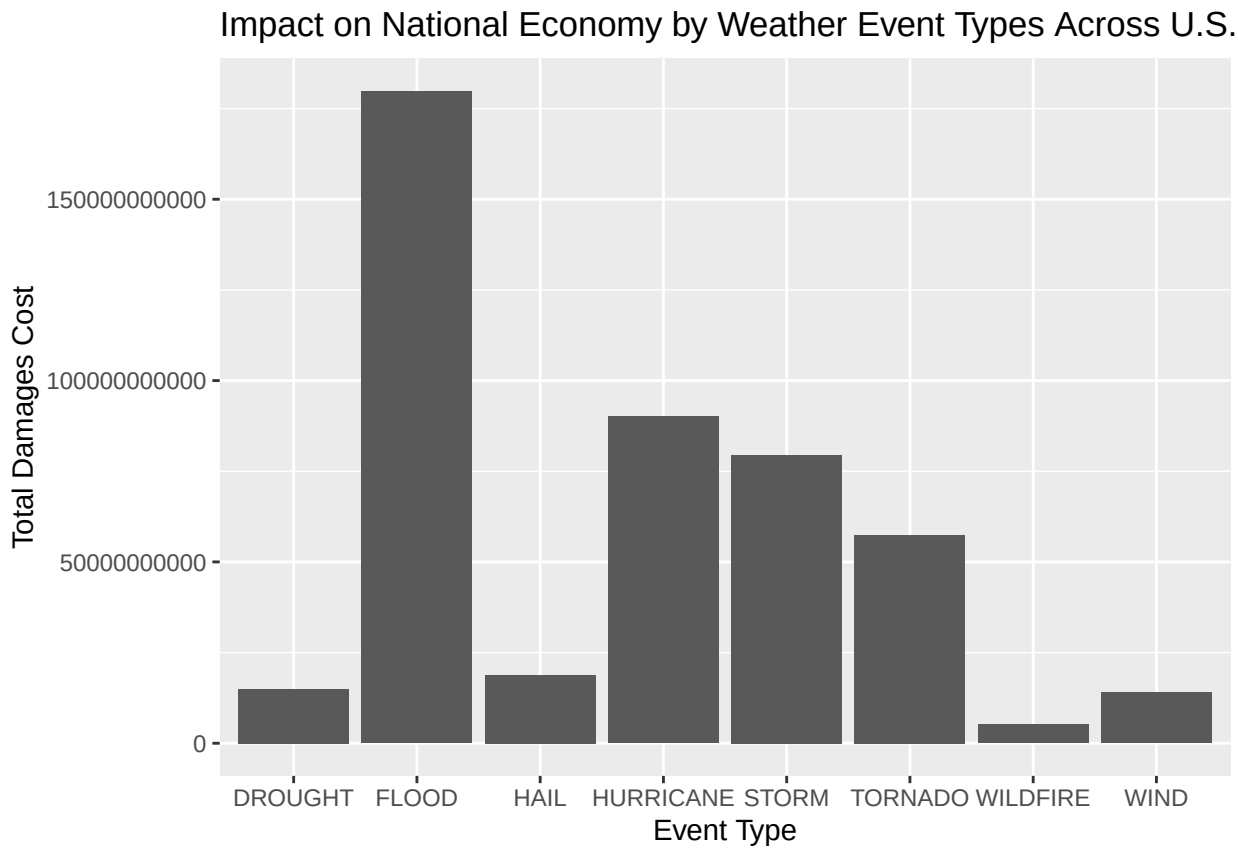
In order to show impacts on economy by different event types, we can make a barplot of number of sum damages from year 1950 to November 2011.

- Merge the data sets related to national economy by event types
- Conduct the plot using ggplot2 system
- Generate the result

```
economyResult <- inner_join(propDMG, cropDMG, by = "New.Event")

economyResult <- economyResult %>% mutate(economyResult, Total=sum.PropCost+sum.CropCost) %>% arrange(-
```

```r
library(ggplot2)
#create the plot
ggplot(economyResult[1:8,], aes(x=New.Event, y=Total))+geom_bar(stat = "identity")+xlab("Event Type")+y
```




```r
head(economyResult,3)
```

```
## # A tibble: 3 x 4
##   New.Event sum.PropCost sum.CropCost        Total
##   <chr>            <dbl>        <dbl>        <dbl>
## 1 FLOOD     167502199413  12266906100 179769105513
## 2 HURRICANE  84656180010   5505292800  90161472810
## 3 STORM      73054022622   6406919600  79460942222
```

The plot combined with the table shows us that the top three weather events with greatest impacts on national economy are **flood, hurricane and storm.** And flood has the obvious greatest influence.