Reproducible Research Week4 project

YH

11/12/2019

Background

Storms and other severe weather events can cause both public health and economic problems for communities and municipalities. Many severe events can result in fatalities, injuries, and property damage, and preventing such outcomes to the extent possible is a key concern. This project involves exploring the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database. This database tracks characteristics of major storms and weather events in the United States, including when and where they occur, as well as estimates of any fatalities, injuries, and property damage.

Synopsis

Here I made the following definition:

1. I define "harmful to health" as the total number of injuries and deaths caused by the event

2. I define "Economic loss" as the total value lost caused by property damage and crop damage

According to the definition, I first compute two variables "Health"(by adding up injuries and deaths) and "Economic"(by adding up property damage and crop damage) Then, I grouped all the event entry into major event categories according to National Weather Service Storm Data Documentation. Finally, I plot barcharts and dotplot to demonstrate that Tornado, Wind and Heat are the three most harmful events to public health, whereas Flood, Hurricane and tide led to the greatest economic lost. In addition, it seems like there is a positive correlation between these two consequence.

Data processing

Setup

```
library(ggplot2)
data<-read.csv("repdata_data_StormData.csv.bz2")
```

Initial loading and exploring data

```
#Explore data
    head(data,2)
```

```
##   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
##   BGN_RANGE BGN_AZI BGN_LOCATI END_DATE END_TIME COUNTY_END COUNTYENDN
## 1         0                                               0         NA
## 2         0                                               0         NA
##   END_RANGE END_AZI END_LOCATI LENGTH WIDTH F MAG FATALITIES INJURIES PROPDMG
## 1         0                      14   100 3   0          0       15    25.0
## 2         0                       2   150 2   0          0        0     2.5
```

```
##   PROPDMGEXP CROPDMG CROPDMGEXP WFO STATEOFFIC ZONENAMES LATITUDE LONGITUDE
## 1          K       0                                         3040      8812
## 2          K       0                                         3042      8755
##   LATITUDE_E LONGITUDE_ REMARKS REFNUM
## 1       3051       8806              1
## 2          0          0              2
```

```
    names(data)
```

```
##  [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"
```

```
#Keep only variables that are relevant to Date/Fatalities/Injuries/Damage
    data1<- data[, c("BGN_DATE","EVTYPE", "FATALITIES",
                     "INJURIES","PROPDMG","PROPDMGEXP","CROPDMG","CROPDMGEXP")]

#See structure of data1
    str(data1)
```

```
## 'data.frame':    902297 obs. of  8 variables:
##  $ BGN_DATE  : Factor w/ 16335 levels "1/1/1966 0:00:00",..: 6523 6523 4242 11116 2224 2224 2260 383
##  $ EVTYPE    : Factor w/ 985 levels "   HIGH SURF ADVISORY",..: 834 834 834 834 834 834 834 834 834 8
##  $ 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 ...
```

Store the cleaned data in a new dataset

```
#Build a new dataset for analysis
     n<-nrow(data1)
     data2<-data.frame(matrix(nrow=n,ncol = 8))
     colnames(data2)<-c("Year","Event",
                        "Deaths","Injury",
                        "Property_damage", "Crop_damage",
                        "Health","Economic")

#Process data and put into new dataset
     data2$Event<-data1$EVTYPE
     data2$Deaths<-data1$FATALITIES
     data2$Injury<-data1$INJURIES

     datetmp<- as.Date(as.character(data1$BGN_DATE),"%m/%d/%Y %H:%M:%S")
     data2$Year<-factor(as.numeric(format(datetmp,"%Y")))
```

```
    #Convert unit of the property damage into real numbers
    summary(data1$PROPDMGEXP)
```

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

```
    data2$Property_damage<-data1$PROPDMG *(
             1 * (data1$PROPDMGEXP =="K" & data1$PROPDMGEXP !="k" &
                  data1$PROPDMGEXP !="M" & data1$PROPDMGEXP !="m" &
                  data1$PROPDMGEXP !="B" & data1$PROPDMGEXP !="b" ) +
             1000 * (data1$PROPDMGEXP =="K" | data1$PROPDMGEXP =="k") +
             1000000 * (data1$PROPDMGEXP =="M" | data1$PROPDMGEXP =="m") +
             1000000000 * (data1$PROPDMGEXP =="B" | data1$PROPDMGEXP =="b")
      )

    #Convert unit of the crop damage into real numbers
    data2$Crop_damage <- data1$CROPDMG *(
             1 * (data1$CROPDMGEXP !="K" & data1$CROPDMGEXP !="k" &
                  data1$CROPDMGEXP !="M" & data1$CROPDMGEXP !="m" &
                  data1$CROPDMGEXP !="B" & data1$CROPDMGEXP !="b" ) +
             1000 * (data1$CROPDMGEXP =="K" | data1$CROPDMGEXP =="k") +
             1000000 * (data1$CROPDMGEXP =="M" | data1$CROPDMGEXP =="m") +
             1000000000 * (data1$CROPDMGEXP =="B" | data1$CROPDMGEXP =="b")
      )
```

Process data

1. I define “harmful to health” as the total number of injuries and deaths caused by the event
2. I define “Economic loss” as the total value lost caused by property damage and crop damage

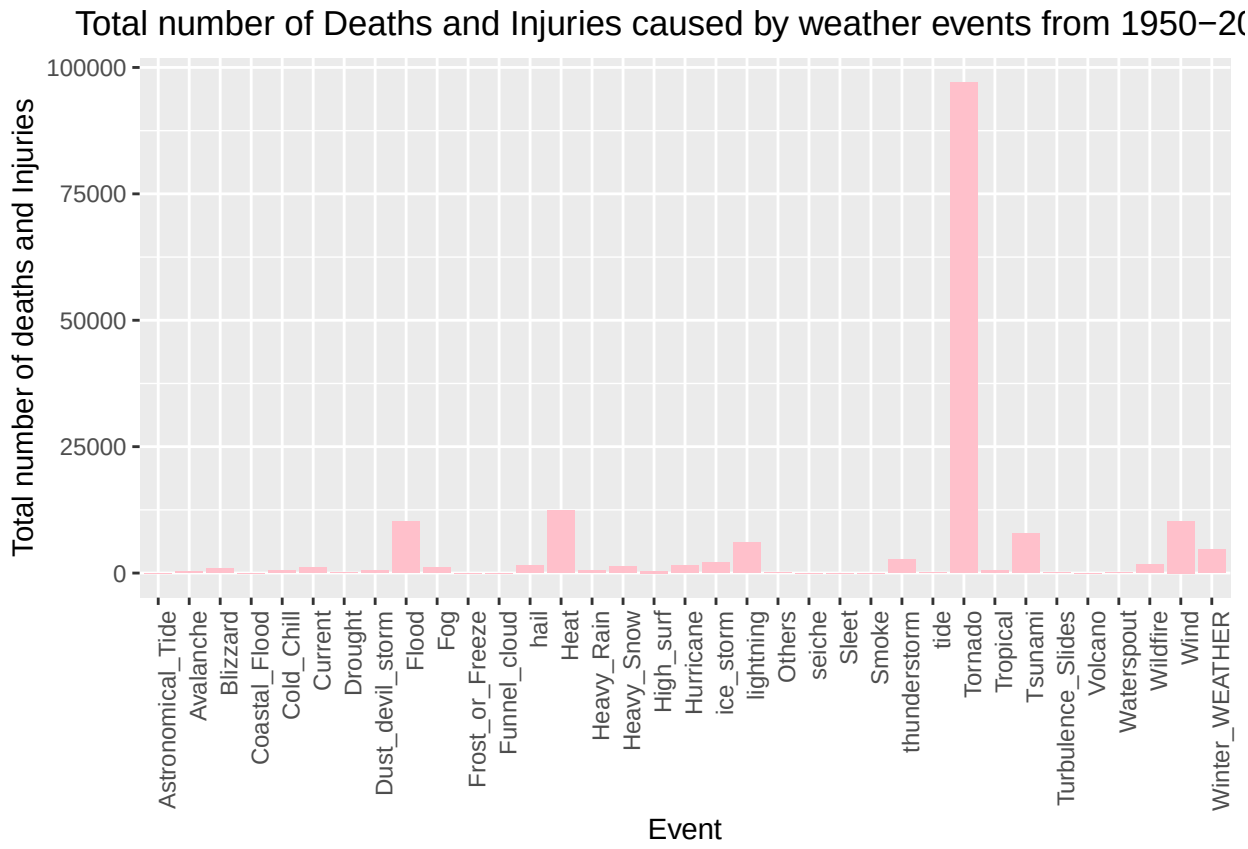
```
#Create new variables for indicating "harmful to health"
    data2$Health <- data2$Deaths + data2$Injury
#Create new variables for indicating "Economic loss"
    data2$Economic <- data2$Property_damage + data2$Crop_damage

#Calculate the total injuries and deaths throught the years by different event
    total_health<-aggregate(Health~Event, data=data2, sum)
#Calculate the total injuries and deaths throught the years by different event
    total_economic<-aggregate(Economic~Event, data=data2, sum)
#Combine the results in a new variable
    total_health_economic<-merge(total_health,total_economic)

#Group weather events according to the storm data event table
    Astronomical_Tide<-grep("Astronomical",total_health_economic$Event,ignore.case = TRUE)
    Avalanche<-grep("Avalanche",total_health_economic$Event,ignore.case = TRUE)
    Blizzard<-grep("Blizzard",total_health_economic$Event,ignore.case = TRUE)
    Coastal_Flood<-grep("Coastal",total_health_economic$Event,ignore.case = TRUE)
    Cold_Chill<-grep("chill|cold|cool|freez|HYPOTHER",total_health_economic$Event,ignore.case = TRUE)
    Fog<-grep("fog",total_health_economic$Event,ignore.case = TRUE)
```

```
Smoke<-grep("smoke",total_health_economic$Event,ignore.case = TRUE)
Dust_devil_storm<-grep("dust",total_health_economic$Event,ignore.case = TRUE)
Heat<-grep("heat|warm|HYPERTHER|HOT|HIGH TEMP",total_health_economic$Event,ignore.case = TRUE)
Flood<-grep("flood|fld|stream|URBAN|STRM",total_health_economic$Event,ignore.case = TRUE)
Frost_or_Freeze<-grep("Frost|Freeze",total_health_economic$Event,ignore.case = TRUE)
Funnel_cloud<-grep("funnel",total_health_economic$Event,ignore.case = TRUE)
hail<-grep("hail",total_health_economic$Event,ignore.case = TRUE)
Heavy_Rain<-grep("rain|rain|rainfall|torrential|SHOWER|unseasonal rain|precipitation|precip",total_
Heavy_Snow<-grep("snow",total_health_economic$Event,ignore.case = TRUE)
High_surf<-grep("surf",total_health_economic$Event,ignore.case = TRUE)
Drought<-grep("dry|drought",total_health_economic$Event,ignore.case = TRUE)
Wind<-grep("Wind|wnd",total_health_economic$Event,ignore.case = TRUE)
Hurricane<-grep("hurricane|typhoon",total_health_economic$Event,ignore.case = TRUE)
ice_storm<-grep("ice * storm",total_health_economic$Event,ignore.case = TRUE)
lightning<-grep("light",total_health_economic$Event,ignore.case = TRUE)
thunderstorm<-grep("thunder|microburst|rainstorm",total_health_economic$Event,ignore.case = TRUE)
Current<-grep("current",total_health_economic$Event,ignore.case = TRUE)
seiche<-grep("seiche",total_health_economic$Event,ignore.case = TRUE)
Sleet<-grep("Sleet",total_health_economic$Event,ignore.case = TRUE)
tide<-grep("storm surge|tIde|HIGH SEA|HIGH SWELL",total_health_economic$Event,ignore.case = TRUE)
Tornado<-grep("tornado",total_health_economic$Event,ignore.case = TRUE)
Tropical<-grep("tropical",total_health_economic$Event,ignore.case = TRUE)
Tsunami<-grep("tsunami|tstm",total_health_economic$Event,ignore.case = TRUE)
Volcano<-grep("volcan",total_health_economic$Event,ignore.case = TRUE)
Waterspout<-grep("water|WET|spout",total_health_economic$Event,ignore.case = TRUE)
Wildfire<-grep("Wild|forest|forrest|fire",total_health_economic$Event,ignore.case = TRUE)
Winter_WEATHER<-grep("WINTER|WINTRY|ICE|glaze",total_health_economic$Event,ignore.case = TRUE)
Turbulence_Slides<-grep("turbulence|slide",total_health_economic$Event,ignore.case = TRUE)

Wind<-Wind[!Wind%in%thunderstorm]
Cold_Chill<-Cold_Chill[!Cold_Chill%in%c(Heavy_Rain,Heavy_Snow,Wind)]
Eventlist<-list(Astronomical_Tide,Avalanche,Blizzard,Coastal_Flood,Cold_Chill,
                Fog,Smoke, Dust_devil_storm,Heat,Flood,
                Frost_or_Freeze,Funnel_cloud,hail,Heavy_Rain,Heavy_Snow, High_surf,
                Hurricane,ice_storm,lightning,thunderstorm,
                Current,seiche,Sleet,tide,Tornado,
                Tropical,Tsunami,Volcano,Waterspout,Wildfire,
                Wind,Drought,Winter_WEATHER,Turbulence_Slides)

Eventnamelist<-c("Astronomical_Tide","Avalanche","Blizzard","Coastal_Flood","Cold_Chill",
                 "Fog","Smoke", "Dust_devil_storm","Heat","Flood",
                 "Frost_or_Freeze","Funnel_cloud","hail","Heavy_Rain","Heavy_Snow", "High_surf",
                 "Hurricane","ice_storm","lightning","thunderstorm",
                 "Current","seiche","Sleet","tide","Tornado",
                 "Tropical","Tsunami","Volcano","Waterspout","Wildfire",
                 "Wind","Drought","Winter_WEATHER","Turbulence_Slides")

#Assign unclassified events into "Others"
Eventlist_n<-unique(as.numeric(unlist(Eventlist)))
Eventlist_n<-sort(Eventlist_n)
Nrow<-as.numeric(1:nrow(total_health_economic))
diff<-Nrow[!Nrow%in%Eventlist_n]
others<-total_health_economic[diff,]
```

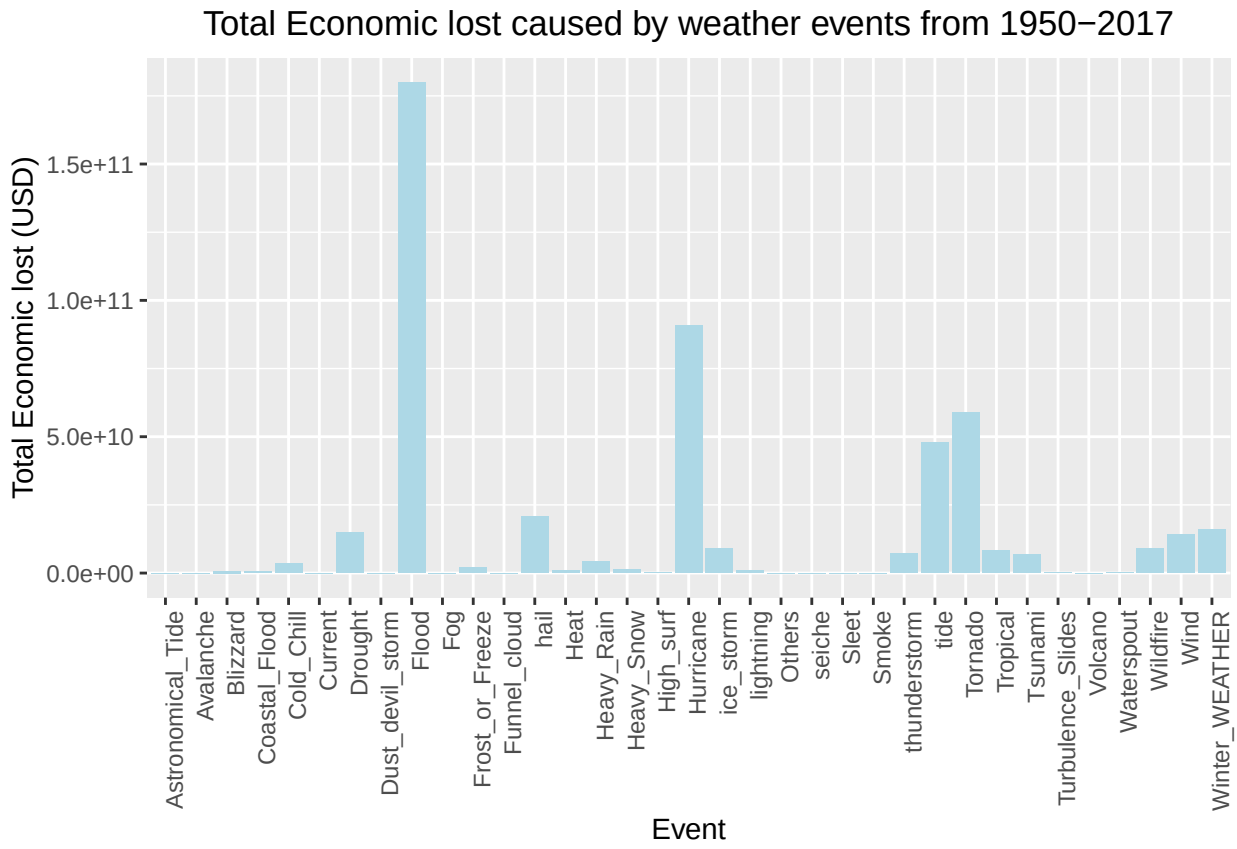
```r
#Create new list
    total<-data.frame(matrix(ncol=3,nrow = length(Eventnamelist)+1))
    colnames(total)<-c("Event","Health","Economic")
    total$Event[1:34]<-Eventnamelist
    total$Event[35]<-"Others"
    temp<-NULL
    for(i in 1:length(Eventnamelist)){
        temp<-total_health_economic[Eventlist[[i]],]
        total$Health[i]<-sum(temp$Health)
        total$Economic[i]<-sum(temp$Economic)
    }
    total$Health[35]<-sum(others$Health)
    total$Economic[35]<-sum(others$Economic)
```

Result

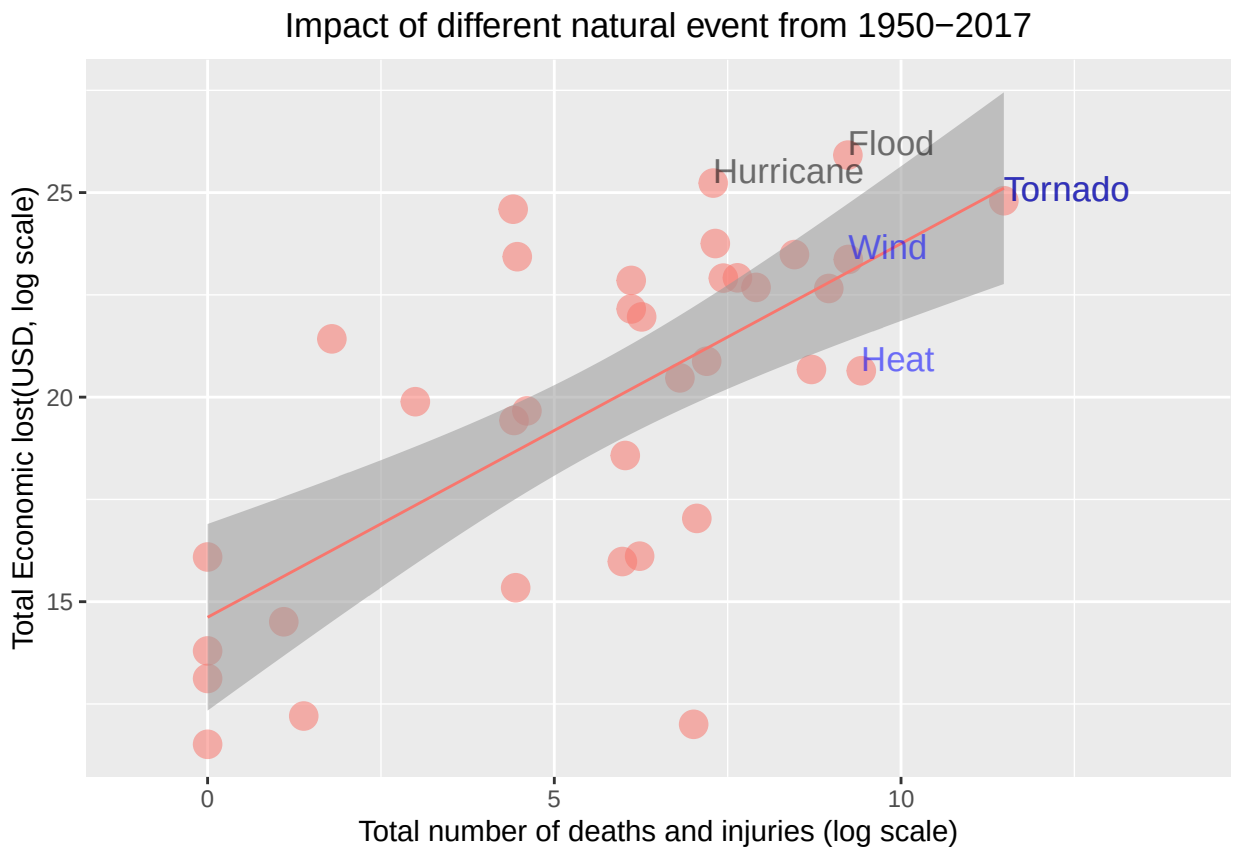
```r
# Plot the total number of death and injuries caused by different events
library(ggplot2)
healthplot<- ggplot(aes(x=Event,y=Health,fill="pink"),data =total) +
             geom_bar(data=total,aes(x=Event,y=Health), stat="identity") +
             theme(axis.text.x = element_text(hjust = 1,angle=90), plot.title = element_text(hjust=0.5)
             ggtitle("Total number of Deaths and Injuries caused by weather events from 1950-2017") +
             ylab("Total number of deaths and Injuries") +
             scale_fill_manual(values="pink")
plot(healthplot)
```




```
# Plot the total number of economic lost caused by different events
economicplot<- ggplot(aes(x=Event,y=Economic,fill="lightblue"), data =total) +
              geom_bar(data=total,aes(x=Event,y=Economic), stat="identity") +
              theme(axis.text.x = element_text(hjust = 1,angle=90), plot.title = element_text(hjust=0.5)
              ggtitle("Total Economic lost caused by weather events from 1950-2017") +
              ylab("Total Economic lost (USD)") +
              scale_fill_manual(values="lightblue")
plot(economicplot)
```




```
#Combine both Health and economic impact
totalplot<-qplot(x=log(Health+1),y=log(Economic+1), data=total,size=2,alpha=0.2,col="pink") +
    theme(legend.position = "none") +
    geom_smooth(method=lm,lwd=0.5) +
    geom_text(aes(label=ifelse(total$Economic>=5.896300e+10,as.character(total$Event),'')),
              hjust=0,vjust=0,col="black") +
    geom_text(aes(label=ifelse(total$Health>10236,as.character(total$Event),'')),
              hjust=0,vjust=0,col="blue") +
    xlim(c(-1,14)) +
    ggtitle("Impact of different natural event from 1950-2017") +
    xlab("Total number of deaths and injuries (log scale)") +
    ylab("Total Economic lost(USD, log scale)")+coord_cartesian(clip = 'off') +
    theme(plot.title = element_text(hjust = 0.5))
plot(totalplot)
```



From these graphs, we can see that Tornado is the most harmful event for public health, followed by Wind and Heat(high temperature).

In terms of economic damage, Flood is on the top of the list, Hurricane and Tornado are the second and third most influential factor.

In addition, the regression model demonstrate the positive correlation between the economic lost and the harm to public health caused by the weather events