Reproducible Research (week4)

Name : Taesoon Kim

Date : Jul-03-2017

Title : Health and economic influences with respect to types of events

Synopsis

There are a variety of severe weather events, and each event influences differently. According to NOAA storm database, I want to know which events are most harmful about public health and have the greatest economic results. For this, I will analyze storm data, and explore in detail to see which events are causing the most damages. Thereafter, if we know what events are coming, we can prepare for prescription. Therefore it can be helpful for our lives. After I analyzed storm data, "TORNADO" is most harmful for population health, and "FLOOD" and "DROUGHT" are the biggest impact on financial damage.

Data Processing

```
# Set the directory
setwd("D:/1-1. R studio/lecture5. reproducible research/week4")

# System change
Sys.setlocale(category="LC_CTYPE",locale="C")
```

```
## [1] "C"
```

```
# Load raw data
raw_data<-read.csv(file="repdata_data_StormData.csv.bz2",header=TRUE)
head(raw_data)
```

```
##   STATE__           BGN_DATE BGN_TIME TIME_ZONE COUNTY COUNTYNAME STATE
## 1       1  4/18/1950 0:00:00     0130       CST     97     MOBILE    AL
## 2       1  4/18/1950 0:00:00     0145       CST      3    BALDWIN    AL
## 3       1  2/20/1951 0:00:00     1600       CST     57    FAYETTE    AL
## 4       1   6/8/1951 0:00:00     0900       CST     89    MADISON    AL
## 5       1 11/15/1951 0:00:00     1500       CST     43    CULLMAN    AL
## 6       1 11/15/1951 0:00:00     2000       CST     77 LAUDERDALE    AL
##    EVTYPE BGN_RANGE BGN_AZI BGN_LOCATI END_DATE END_TIME COUNTY_END
## 1 TORNADO         0                                               0
## 2 TORNADO         0                                               0
## 3 TORNADO         0                                               0
## 4 TORNADO         0                                               0
## 5 TORNADO         0                                               0
## 6 TORNADO         0                                               0
##   COUNTYENDN END_RANGE END_AZI END_LOCATI LENGTH WIDTH F MAG FATALITIES
## 1         NA         0                      14.0   100 3   0          0
## 2         NA         0                       2.0   150 2   0          0
## 3         NA         0                       0.1   123 2   0          0
## 4         NA         0                       0.0   100 2   0          0
## 5         NA         0                       0.0   150 2   0          0
```

```
## 6         NA         0                      1.5   177 2   0          0
##   INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP WFO STATEOFFIC ZONENAMES
## 1       15    25.0          K       0
## 2        0     2.5          K       0
## 3        2    25.0          K       0
## 4        2     2.5          K       0
## 5        2     2.5          K       0
## 6        6     2.5          K       0
##   LATITUDE LONGITUDE LATITUDE_E LONGITUDE_ REMARKS REFNUM
## 1     3040      8812       3051       8806              1
## 2     3042      8755          0          0              2
## 3     3340      8742          0          0              3
## 4     3458      8626          0          0              4
## 5     3412      8642          0          0              5
## 6     3450      8748          0          0              6
```

In CSV file, there are 37 columns and 902,297 rows. As I use head() function, I can check the column names, and which data is in raw file.

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

- I already saw data variable, "FATALITIES" and "INJURIES" data are influenced by events

```
# Identify the EVTYPE labels
events<-unique(raw_data$EVTYPE)

# How "Fatalities" are influenced
# I align descending order, and select 10 rows
fatalities<-aggregate(FATALITIES~EVTYPE,raw_data,sum)
fatal_order<-fatalities[order(-fatalities$FATALITIES),]
fatal_order_head<-head(fatal_order,10)
fatal_order_head
```

```
##             EVTYPE FATALITIES
## 834        TORNADO       5633
## 130 EXCESSIVE HEAT       1903
## 153    FLASH FLOOD        978
## 275           HEAT        937
## 464      LIGHTNING        816
## 856      TSTM WIND        504
## 170          FLOOD        470
## 585    RIP CURRENT        368
## 359      HIGH WIND        248
## 19       AVALANCHE        224
```

```
# How "Injuries" are influenced
# I align descending order, and select 10 rows
injuries<-aggregate(INJURIES~EVTYPE,raw_data,sum)
inj_order<-injuries[order(-injuries$INJURIES),]
inj_order_head<-head(inj_order,10)
inj_order_head
```

```
##               EVTYPE INJURIES
## 834          TORNADO    91346
## 856        TSTM WIND     6957
```

```
## 170             FLOOD     6789
## 130    EXCESSIVE HEAT     6525
## 464         LIGHTNING     5230
## 275              HEAT     2100
## 427         ICE STORM     1975
## 153       FLASH FLOOD     1777
## 760 THUNDERSTORM WIND     1488
## 244              HAIL     1361
```

2. Across the United States, which types of events have the greatest economic consequences?

- I already saw data variable, "Property damage" and "Crop damage" data are influenced by events

```
# Identify the Property damage labels
prop_dmg_exp<-unique(raw_data$PROPDMGEXP)
prop_dmg_exp
```

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

```
# Property damage exp has 19 levels, and I allocate the number
raw_data$PROP[raw_data$PROPDMGEXP=="-"]<--1
raw_data$PROP[raw_data$PROPDMGEXP=="?"]<-0
raw_data$PROP[raw_data$PROPDMGEXP=="+"|raw_data$PROPDMGEXP==""]<-+1
for(i in 0:8){
  raw_data$PROP[raw_data$PROPDMGEXP==i]<-(10^i)
}
raw_data$PROP[raw_data$PROPDMGEXP=="B"]<-(10^9)
raw_data$PROP[raw_data$PROPDMGEXP=="h"|raw_data$PROPDMGEXP=="H"]<-(10^2)
raw_data$PROP[raw_data$PROPDMGEXP=="K"]<-(10^3)
raw_data$PROP[raw_data$PROPDMGEXP=="m"|raw_data$PROPDMGEXP=="M"]<-(10^6)

# I will calculate the property damage, multiplying "PROPDMG" and "PROPDMGEXP"
raw_data$PROPVAL=raw_data$PROPDMG*raw_data$PROP

# How "Property damage" is influenced
# I align descending order, and select 10 rows
prop_damage<-aggregate(PROPVAL~EVTYPE,raw_data,sum)
prop_order<-prop_damage[order(-prop_damage$PROPVAL),]
prop_order_head<-head(prop_order,10)
prop_order_head
```

```
##                EVTYPE      PROPVAL
## 170             FLOOD 144657709807
## 411 HURRICANE/TYPHOON  69305840000
## 834           TORNADO  56947380677
## 670       STORM SURGE  43323536000
## 153       FLASH FLOOD  16822673979
## 244              HAIL  15735267513
## 402         HURRICANE  11868319010
## 848    TROPICAL STORM   7703890550
## 972      WINTER STORM   6688497251
## 359         HIGH WIND   5270046265
```

```
# Identify the crop damage labels
crop_dmg_exp<-unique(raw_data$CROPDMGEXP)
crop_dmg_exp
```

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

```
# Crop damage exp has 9 levels, and I allocate the number
raw_data$CROP[raw_data$CROPDMGEXP=="?"]<-0
for(i in 0:8){
  raw_data$CROP[raw_data$CROPDMGEXP==i]<-(10^i)
}
raw_data$CROP[raw_data$CROPDMGEXP=="B"]<-(10^9)
raw_data$CROP[raw_data$CROPDMGEXP=="K"]<-(10^3)
raw_data$CROP[raw_data$CROPDMGEXP=="m"|raw_data$CROPDMGEXP=="M"]<-(10^6)

# I will calculate the crop damage, multiplying "CROPDMG" and "CROPDMGEXP"
raw_data$CROPVAL=raw_data$CROPDMG*raw_data$CROP

# How "Crop damage" is influenced
# I align descending order, and select 10 rows
crop_damage<-aggregate(CROPVAL~EVTYPE,raw_data,sum)
crop_order<-crop_damage[order(-crop_damage$CROPVAL),]
crop_order_head<-head(crop_order,10)
crop_order_head
```

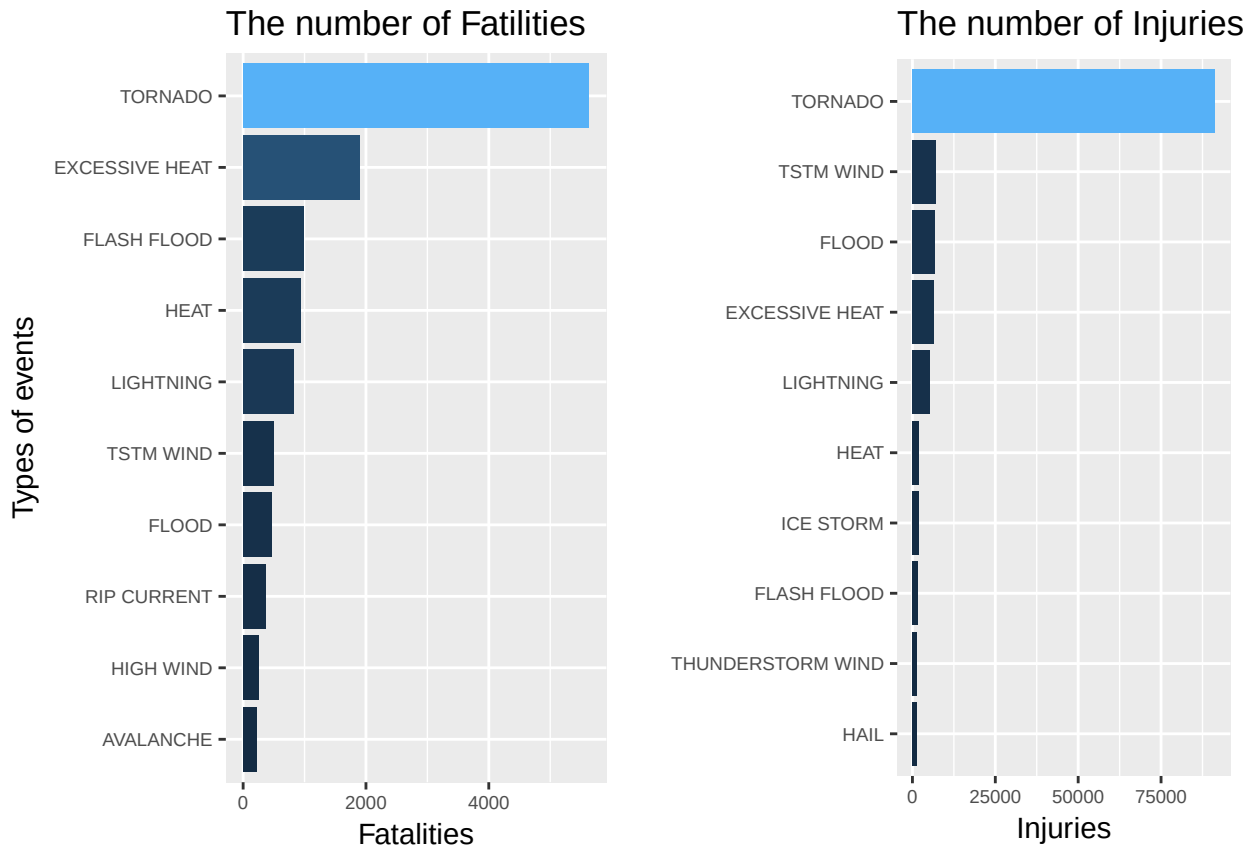
```
##               EVTYPE     CROPVAL
## 16           DROUGHT 13972566000
## 35             FLOOD  5661968450
## 99       RIVER FLOOD  5029459000
## 86         ICE STORM  5022113500
## 53              HAIL  3025537470
## 78         HURRICANE  2741910000
## 83 HURRICANE/TYPHOON  2607872800
## 30       FLASH FLOOD  1421317100
## 26      EXTREME COLD  1292973000
## 47      FROST/FREEZE  1094086000
```

Results

```
# plot graph
library(ggplot2)
library(gridExtra)
g1<-ggplot(fatal_order_head,aes(x=reorder(EVTYPE,FATALITIES),y=FATALITIES,fill=FATALITIES))
g1<-g1+geom_bar(stat="identity")+coord_flip()
g1<-g1+labs(title="The number of Fatilities",x="Types of events",y="Fatalities")
g1<-g1+theme(legend.position = "none",axis.text=element_text(size=7))

g2<-ggplot(inj_order_head,aes(x=reorder(EVTYPE,INJURIES),y=INJURIES,fill=INJURIES))
g2<-g2+geom_bar(stat="identity")+coord_flip()
g2<-g2+labs(title="The number of Injuries",x="",y="Injuries")
g2<-g2+theme(legend.position = "none",axis.text=element_text(size=7))
```

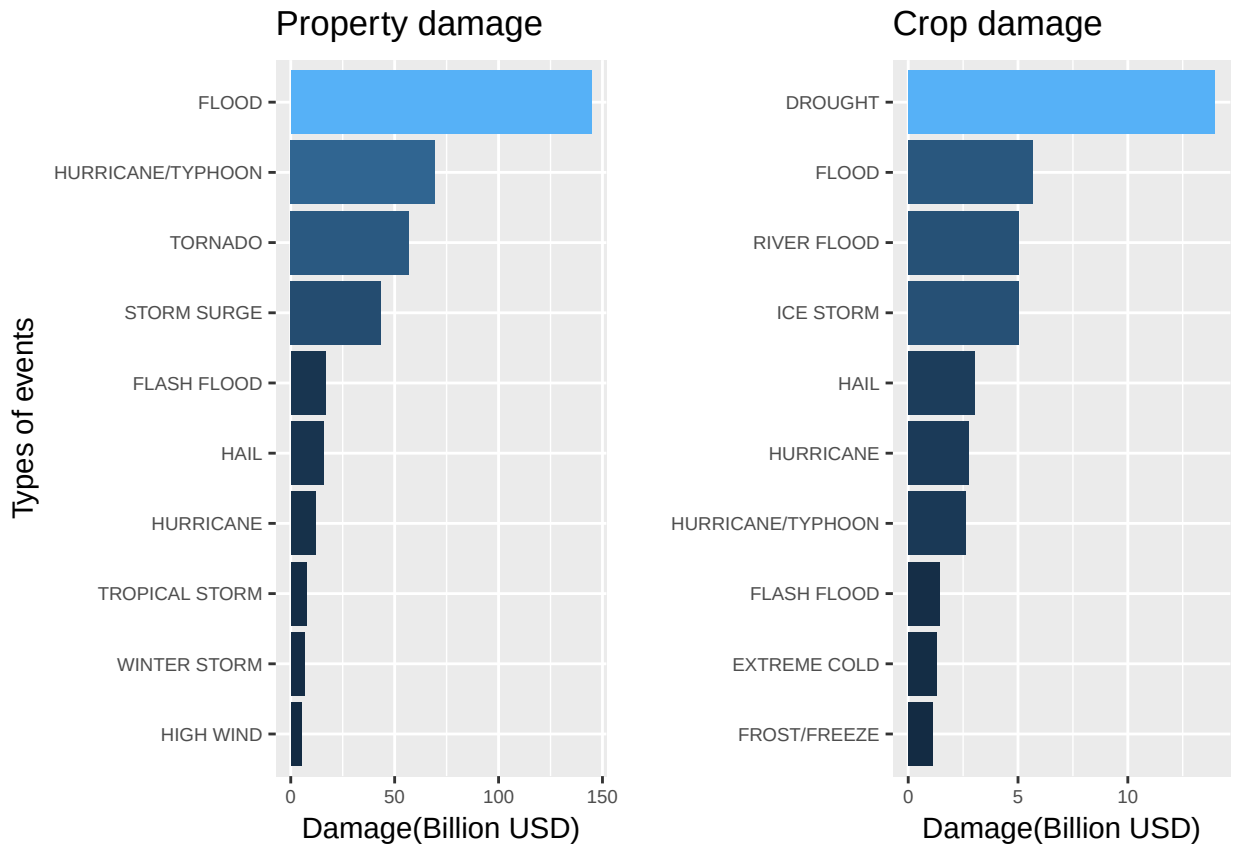
```
grid.arrange(g1, g2, ncol=2)
```




```
# plot graph
prop_order_head$PROPVAL<-prop_order_head$PROPVAL/(10^9)    # change the unit, -> Billion
g3<-ggplot(prop_order_head,aes(x=reorder(EVTYPE,PROPVAL),y=PROPVAL,fill=PROPVAL))
g3<-g3+geom_bar(stat="identity")+coord_flip()
g3<-g3+labs(title="Property damage",x="Types of events",y="Damage(Billion USD)")
g3<-g3+theme(legend.position ="none",axis.text=element_text(size=7))

crop_order_head$CROPVAL<-crop_order_head$CROPVAL/(10^9)
g4<-ggplot(crop_order_head,aes(x=reorder(EVTYPE,CROPVAL),y=CROPVAL,fill=CROPVAL))
g4<-g4+geom_bar(stat="identity")+coord_flip()
g4<-g4+labs(title="Crop damage",x="",y="Damage(Billion USD)")
g4<-g4+theme(legend.position="none",axis.text=element_text(size=7))

grid.arrange(g3, g4, ncol=2)
```



First, above the graph, "fatalities" and "injuries" are most influenced by "Tornado". Second, "FLOOD" and "DROUGHT" have the greatest economic consequences.