Course Project 2, Storms

Miguel Porro

October 22, 2018

Reproducible Research Course Project

Synopsis

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. From these data, we found that, **TORNADO** is the event that most harmful with respect to population health, while **FLOOD** is the event that most harmful with respect to population health.

Loading Raw Data

From the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database, we obtained the data in the form of a comma-seperated-value file compressed via the bzip2 algorithm to reduce the size.

```
storm_data <- read.csv("C:/Users/migue/Documents/Coursera/repdata.csv",
header = TRUE, sep = ",", na.string = "")
```

After loading, we read a few rows in this dataset

```
dim(storm_data)

## [1] 902297     37

head(storm_data[, 1:13])

##   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
## 1 TORNADO         0    <NA>       <NA>     <NA>     <NA>
## 2 TORNADO         0    <NA>       <NA>     <NA>     <NA>
## 3 TORNADO         0    <NA>       <NA>     <NA>     <NA>
## 4 TORNADO         0    <NA>       <NA>     <NA>     <NA>
## 5 TORNADO         0    <NA>       <NA>     <NA>     <NA>
## 6 TORNADO         0    <NA>       <NA>     <NA>     <NA>
```

Then check the data variables and its characteristics

```
str(storm_data)
```

```
## '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 3980 3980 ...
##  $ BGN_TIME  : Factor w/ 3608 levels "00:00:00 AM",..: 272 287 2705
1683 2584 3186 242 1683 3186 3186 ...
##  $ 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/ 29600 levels "5NM E OF MACKINAC BRIDGE TO
PRESQUE ISLE LT MI",..: 13512 1872 4597 10591 4371 10093 1972 23872 24417
4597 ...
##  $ 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
834 834 834 834 834 834 834 834 834 ...
##  $ BGN_RANGE : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BGN_AZI   : Factor w/ 34 levels "  N"," NW","E",..: NA NA NA NA NA
NA NA NA NA NA ...
##  $ BGN_LOCATI: Factor w/ 54428 levels "- 1 N Albion",..: NA NA NA NA
NA NA NA NA NA NA ...
##  $ END_DATE  : Factor w/ 6662 levels "1/1/1993 0:00:00",..: NA NA NA
NA NA NA NA NA NA NA ...
##  $ END_TIME  : Factor w/ 3646 levels " 0900CST"," 200CST",..: NA NA NA
NA NA NA NA NA NA NA ...
##  $ 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/ 23 levels "E","ENE","ESE",..: NA NA NA NA NA
NA NA NA NA NA ...
##  $ END_LOCATI: Factor w/ 34505 levels "- .5 NNW","- 11 ESE Jay",..: NA
NA NA NA NA NA NA NA NA NA ...
##  $ 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/ 18 levels "-","?","+","0",..: 16 16 16 16 16
16 16 16 16 16 ...
##  $ CROPDMG   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CROPDMGEXP: Factor w/ 8 levels "?","0","2","B",..: NA NA NA NA NA
NA NA NA NA NA ...
##  $ WFO       : Factor w/ 541 levels " CI","$AC","$AG",..: NA NA NA NA
NA NA NA NA NA NA ...
##  $ STATEOFFIC: Factor w/ 249 levels "ALABAMA, Central",..: NA NA NA NA
NA NA NA NA NA NA ...
##  $ ZONENAMES : Factor w/ 25111 levels "
"| __truncated__,..: NA NA NA NA NA NA NA NA NA NA ...
##  $ 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/ 436773 levels "-2 at Deer Park\n",..: NA NA
NA NA NA NA NA NA NA NA ...
##  $ REFNUM    : num  1 2 3 4 5 6 7 8 9 10 ...
```

Data Processing

Which Type of Events are Most Harmful with Respect to Population Health

We will concentrate on two particular variables, **FATALITIES** and **INJUREIS**. So first let’s group the data based on the type of the event **EVTYPE**.

```
data_INJ <- aggregate(storm_data["INJURIES"], list(EVTYPE =
storm_data$EVTYPE), sum)
data_FAT <- aggregate(storm_data["FATALITIES"], list(EVTYPE =
storm_data$EVTYPE), sum)
data_PH <- merge(data_INJ, data_FAT, by = "EVTYPE", all = TRUE)
summary(data_PH)

##                       EVTYPE       INJURIES         FATALITIES
##     HIGH SURF ADVISORY:  1   Min.   :    0.0   Min.   :   0.00
##   COASTAL FLOOD       :  1   1st Qu.:    0.0   1st Qu.:   0.00
##   FLASH FLOOD         :  1   Median :    0.0   Median :   0.00
##   LIGHTNING           :  1   Mean   :  142.7   Mean   :  15.38
##   TSTM WIND           :  1   3rd Qu.:    0.0   3rd Qu.:   0.00
##   TSTM WIND (G45)     :  1   Max.   :91346.0   Max.   :5633.00
##  (Other)              :979
```

It is shown that there are a total of 979 types of weather events. a scatterplot was made to measure which events has the more impact on both, Injuries and Fatalities.

- The injuries average number is 142.7
- The fatalities average reached a number of 15.38

So in order to make the plot easy to read, it was chosen the point which contains injuries number larger then the mean.

Which Type of Events have the Greatest Economic Consequences

To address this question, it was selected the **PROPDMG**, **PROPDMGEXP**, **CROPDMG**, **CROPDMGEXP** variables. These 4 variables, given in numerical values, represents the magnitude of the damage caused to the property. However, **PROPDMGEXP** and **CROPDMGEXP** represents the multiples in thousands *K* amd millions *M*, for the corresponding value for crop damage and property damage. Therefore, we just choose the highest multipler **M** for our analysis.

```
data.sub <- subset(storm_data, select = c(EVTYPE, PROPDMG, PROPDMGEXP,
CROPDMG, CROPDMGEXP))
data.sub1 <- subset(data.sub, data.sub$PROPDMGEXP %in% "M")
data.sub2 <- subset(data.sub1, data.sub1$CROPDMGEXP %in% "M")
head(data.sub2)

##                 EVTYPE PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP
## 187581 HURRICANE ERIN    25.0          M       1          M
## 187583 HURRICANE OPAL    48.0          M       4          M
## 188204       FLOODING    50.0          M       5          M
## 188205     HEAVY RAIN    50.0          M       5          M
## 191345   WINTER STORM     5.0          M       5          M
## 192339     HIGH WINDS     5.5          M       7          M
```

First I selected all the value that **PROPDMGEXP** and **CROPDMGEXP** are equals to *B*

```
data_PRO <- aggregate(data.sub2["PROPDMG"], list(EVTYPE =
data.sub2$EVTYPE), sum)
data_CRO <- aggregate(data.sub2["CROPDMG"], list(EVTYPE =
data.sub2$EVTYPE), sum)
data_ECO <- merge(data_PRO, data_CRO, by = "EVTYPE", all = TRUE)
```

Then it was merged the needed data together to make a plot.

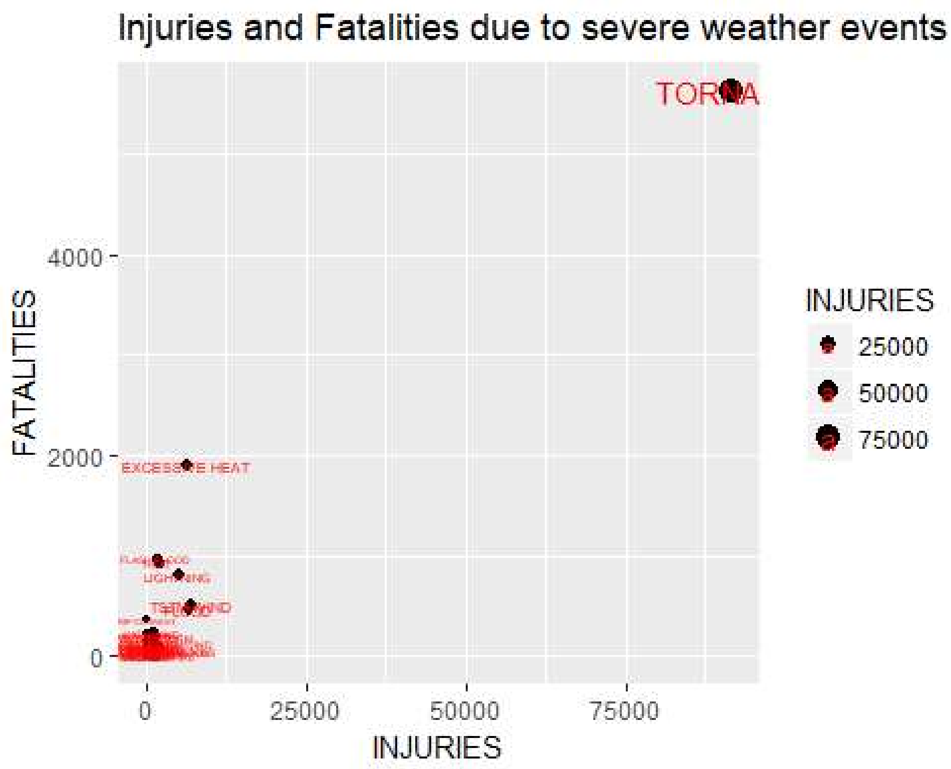
Results

Injuries and Fatalities due to severe weather events

```
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.4.4

g <- ggplot(data_PH[data_PH$INJURIES > 142.7, ], aes(INJURIES,
FATALITIES, label = EVTYPE))
g + geom_point(aes(size = INJURIES)) + geom_text(aes(size = INJURIES),
colour = "red") + scale_size(range = c(1, 4)) + labs(title = "Injuries
and Fatalities due to severe weather events")
```



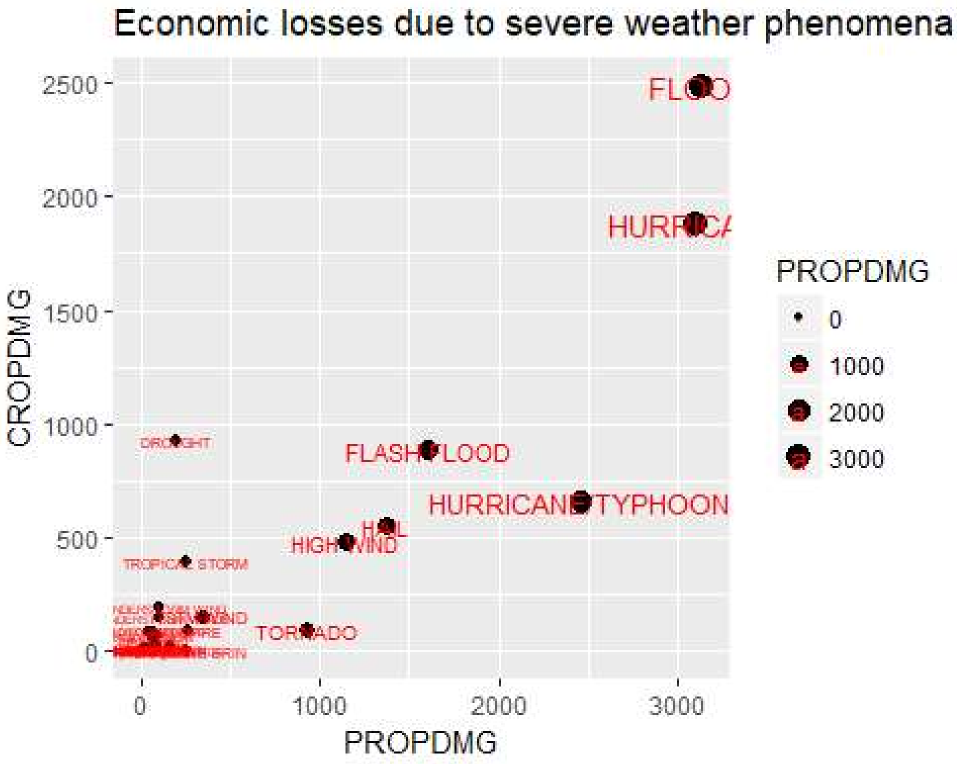
According to the plot, **TORNADO** is the event that most harmful with respect to population health. To see it clearly, I choosed top six event, and listed the Injuries number and Fatalities ruined number below. The ording rule is that Injuries first then Fatalities.

```
head(data_PH[order(data_PH$INJURIES, data_PH$FATALITIES, decreasing =
TRUE), ])
```

```
##              EVTYPE INJURIES FATALITIES
## 834         TORNADO    91346       5633
## 856       TSTM WIND     6957        504
## 170           FLOOD     6789        470
## 130 EXCESSIVE HEAT     6525       1903
## 464       LIGHTNING     5230        816
## 275            HEAT     2100        937
```

Economic losses due to severe weather phenomena

```
library(ggplot2)
g <- ggplot(data_ECO, aes(PROPDMG, CROPDMG, label = EVTYPE))
g + geom_point(aes(size = PROPDMG)) + geom_text(aes(size = PROPDMG),
colour = "red") + scale_size(range = c(1, 4)) + labs(title = "Economic
losses due to severe weather phenomena")
```



According to the plot, **FLOOD** is the event that most harmful with respect to population health. To see it clearly, I choosed top six event, and listed the Injuries number and Fatalities ruined number below. The ording rule is that Injuries first then Fatalities.

```
head(data_ECO[order(data_ECO$PROPDMG, data_ECO$CROPDMG, decreasing =
TRUE), ])
```

```
##                EVTYPE PROPDMG CROPDMG
## 5               FLOOD 3136.64 2487.21
## 17          HURRICANE 3105.87 1879.31
## 20 HURRICANE/TYPHOON 2460.75  656.64
## 3         FLASH FLOOD 1614.40  880.56
## 9                HAIL 1372.87  550.15
## 14          HIGH WIND 1150.09  481.50
```