Reproducible-Research—Course-Project-2

Zanin Pavel

February 29, 2016

Synopsis

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.

Data

The data for this assignment come in the form of a comma-separated-value file compressed via the bzip2 algorithm to reduce its size. You can find file in course work project folder class .

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 see
- National Climatic Data Center Storm Events FAQ see

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.

Assignment

The basic goal of this assignment is to explore the NOAA Storm Database and answer some basic questions about severe weather events. You must use the database to answer the questions below and show the code for your entire analysis. Your analysis can consist of tables, figures, or other summaries. You may use any R package you want to support your analysis.

Questions

Your data analysis must address the following questions:

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?

Consider writing your report as if it were to be read by a government or municipal manager who might be responsible for preparing for severe weather events and will need to prioritize resources for different types of events. However, there is no need to make any specific recommendations in your report.

1. Data Processing

Loading the data.

```
storm.data = read.csv(bzfile("repdata-data-StormData.csv.bz2"), header = TRUE)
```

Loaded data's summary:

```
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
##  $ 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 "","- 1 N Albion",..: 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 "","- .5 NNW",..: 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","$AC",..: 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 "","-2 at Deer Park\n",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ REFNUM    : num  1 2 3 4 5 6 7 8 9 10 ...
```

Loading required packages

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.3
```

```r
library(plyr)
```

```r
storm.data$PROPMULT <- 1
storm.data$PROPMULT[storm.data$PROPDMGEXP =="H"] <- 100
storm.data$PROPMULT[storm.data$PROPDMGEXP =="K"] <- 1000
storm.data$PROPMULT[storm.data$PROPDMGEXP =="M"] <- 1000000
storm.data$PROPMULT[storm.data$PROPDMGEXP =="B"] <- 1000000000

storm.data$CROPMULT <- 1
storm.data$CROPMULT[storm.data$CROPDMGEXP =="H"] <- 100
storm.data$CROPMULT[storm.data$CROPDMGEXP =="K"] <- 1000
storm.data$CROPMULT[storm.data$CROPDMGEXP =="M"] <- 1000000
storm.data$CROPMULT[storm.data$CROPDMGEXP =="B"] <- 1000000000
```

Then summarizing the selected data.

```r
shortStormData <- ddply(.data = storm.data, .variables = .(EVTYPE),
                        fatalities = sum(FATALITIES),
                        injuries = sum(INJURIES),
                        property_damage = sum(PROPDMG * PROPMULT),
                        crop_damage = sum(CROPDMG * CROPMULT),
                        summarize)
```

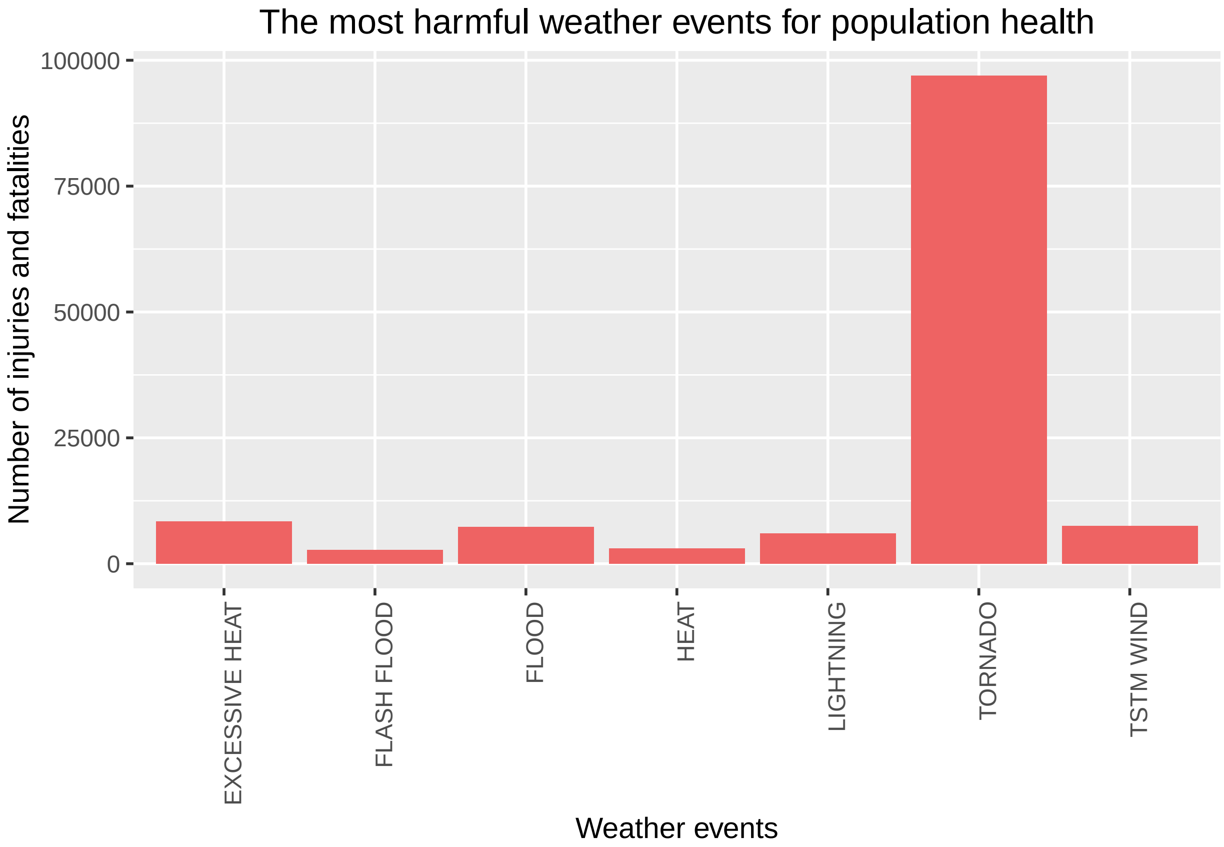
```r
harmfulForHealthEvents <- arrange(shortStormData, desc(fatalities + injuries))
```

```r
economicConsequences <- arrange(shortStormData, desc(property_damage + crop_damage))
```

2. Results

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

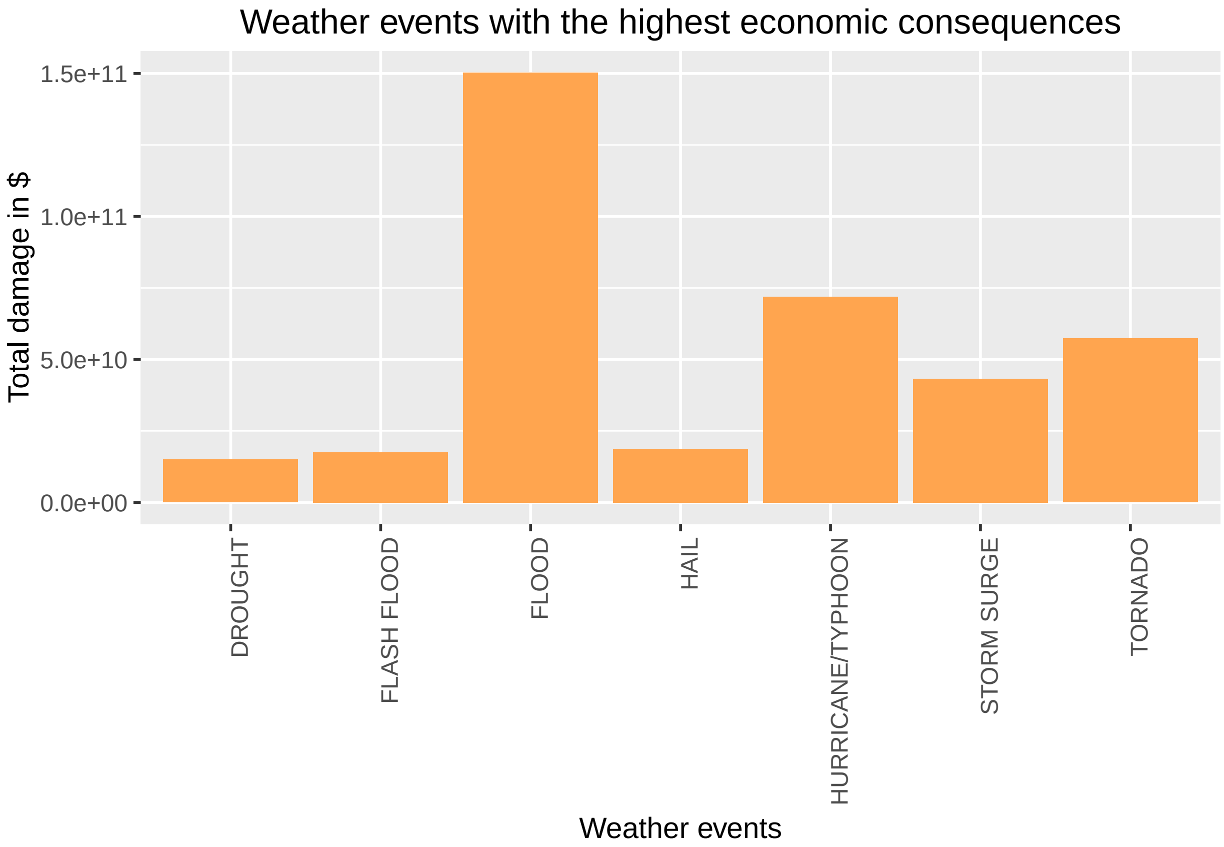
```r
ggplot(data = head(harmfulForHealthEvents, 7),
       aes(x = factor(EVTYPE),
           y = (fatalities + injuries),
           fill = EVTYPE)) +
  geom_bar(stat="identity",
           fill="#EE6363") +
  theme(axis.text.x = element_text(angle = 90,
                                   hjust = 1)) +
  ggtitle("The most harmful weather events for population health") +
  xlab("Weather events") +
  ylab("Number of injuries and fatalities")
```



Answer: The most harmful weather events for population health is Tornado with 96979 victims.

2.2 Question 2 - Across the United States, which types of events have the greatest economic consequences?

```
ggplot(data = head(economicConsequences, 7),
       aes(x = factor(EVTYPE),
           y = property_damage + crop_damage,
           fill = EVTYPE)) +
  geom_bar(stat="identity",
           fill="#FFA54F") +
  theme(axis.text.x = element_text(angle = 90,
                                   hjust = 1)) +
  ggtitle("Weather events with the highest economic consequences") +
  xlab("Weather events") +
  ylab("Total damage in $")
```



Answer: Event with the highest economic consequences is Flood with 5661968450 $ damage.