Reproducible Research - Course Project 2

Pavit Masson

October 22, 2018

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.

The basic goal of this assignment is to explore the NOAA Storm Database and answer some questions, such as:

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

Data Processing

The necessary packages and the data from the storm database used for this analysis were loaded as follows:

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
url <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
download.file(url, "StormData.csv")
data <- read.csv("StormData.csv")
data <- tbl_df(data)
```

For the first question, we will examine the number of fatalities and injuries per event type (EVTYPE).

To get an idea of the data, we'll look at the data in three different ways: by most fatalities (fatal_data), by most injuries (inj_data), and by most fatalities plus injuries (q1_data).

```
fatal_data <- data %>% group_by(EVTYPE) %>% summarise(tot_fatal = sum(FATALITIES)) %>% arrange(desc(tot_

inj_data <- data %>% group_by(EVTYPE) %>% summarise(tot_inj = sum(INJURIES)) %>% arrange(desc(tot_inj))
```

```r
q1_data <- data %>% group_by(EVTYPE) %>% summarise(total = sum(FATALITIES) + sum(INJURIES)) %>% arrange
```

For Question 2, the variables we will look at are PROPDMG, PROPDMGEXP, CROPDMG, CROPDMGEXP. According to the guide for the data, we'll use the EXP variables to apply multipliers to PROPDMG and CROPDMG, depending on the values M, B, and K. We'll assume the other values are errors, since they are not mentioned in the guide.

```r
q2_data <- data
q2_data$PROPDMGEXP <- toupper(as.character(q2_data$PROPDMGEXP))
q2_data$CROPDMGEXP <- as.character(q2_data$CROPDMGEXP)

# Convert PROPDMGEXP and CROPDMGEXP
q2_data[q2_data$PROPDMGEXP == "K", "PROPDMGEXP"] <- 1000
q2_data[q2_data$PROPDMGEXP == "M", "PROPDMGEXP"] <- 1000000
q2_data[q2_data$PROPDMGEXP == "B", "PROPDMGEXP"] <- 1000000000
q2_data$PROPDMGEXP <- as.numeric(q2_data$PROPDMGEXP)
```

```
## Warning: NAs introduced by coercion
```

```r
q2_data$PROPDMGEXP <- q2_data$PROPDMGEXP %>% replace(., is.na(.), 0)

q2_data[q2_data$CROPDMGEXP %in% c("K","k"), "CROPDMGEXP"] <- 1000
q2_data[q2_data$CROPDMGEXP %in% c("M","m"), "CROPDMGEXP"] <- 1000000
q2_data[q2_data$CROPDMGEXP %in% c("B","b"), "CROPDMGEXP"] <- 1000000000

q2_data$CROPDMGEXP <- as.numeric(q2_data$CROPDMGEXP)
```

```
## Warning: NAs introduced by coercion
```

```r
q2_data$CROPDMGEXP <- q2_data$CROPDMGEXP %>% replace(., is.na(.), 0)

q2_data$PropertyDamage <- as.numeric(q2_data$PROPDMG * q2_data$PROPDMGEXP)
q2_data$CropDamage <- as.numeric(q2_data$CROPDMG * q2_data$CROPDMGEXP)
```

Similar to question 1, we'll look at the data in three different ways: by most property damage (prop_data), by most crop damage (crop_data), and by the combined total of property and crop damage (q2_data).

```r
prop_data <- q2_data %>% group_by(EVTYPE) %>% summarize(TotalPropertyDamage = sum(PropertyDamage)) %>% a

crop_data <- q2_data %>% group_by(EVTYPE) %>% summarize(TotalCropDamage = sum(CropDamage)) %>% arrange(

q2_data <- q2_data %>% group_by(EVTYPE) %>% summarize(Total = sum(PropertyDamage) + sum(CropDamage)) %>%
```

Results

Question 1

When we look at the data arranged by highest fatalities, we can see Tornado is first by a big margin.

```r
fatal_data
```

```
## # A tibble: 985 x 2
##    EVTYPE         tot_fatal
##    <fct>              <dbl>
##  1 TORNADO             5633
##  2 EXCESSIVE HEAT      1903
```

```
##   3 FLASH FLOOD          978
##   4 HEAT                 937
##   5 LIGHTNING            816
##   6 TSTM WIND            504
##   7 FLOOD                470
##   8 RIP CURRENT          368
##   9 HIGH WIND            248
##  10 AVALANCHE            224
## # ... with 975 more rows
```

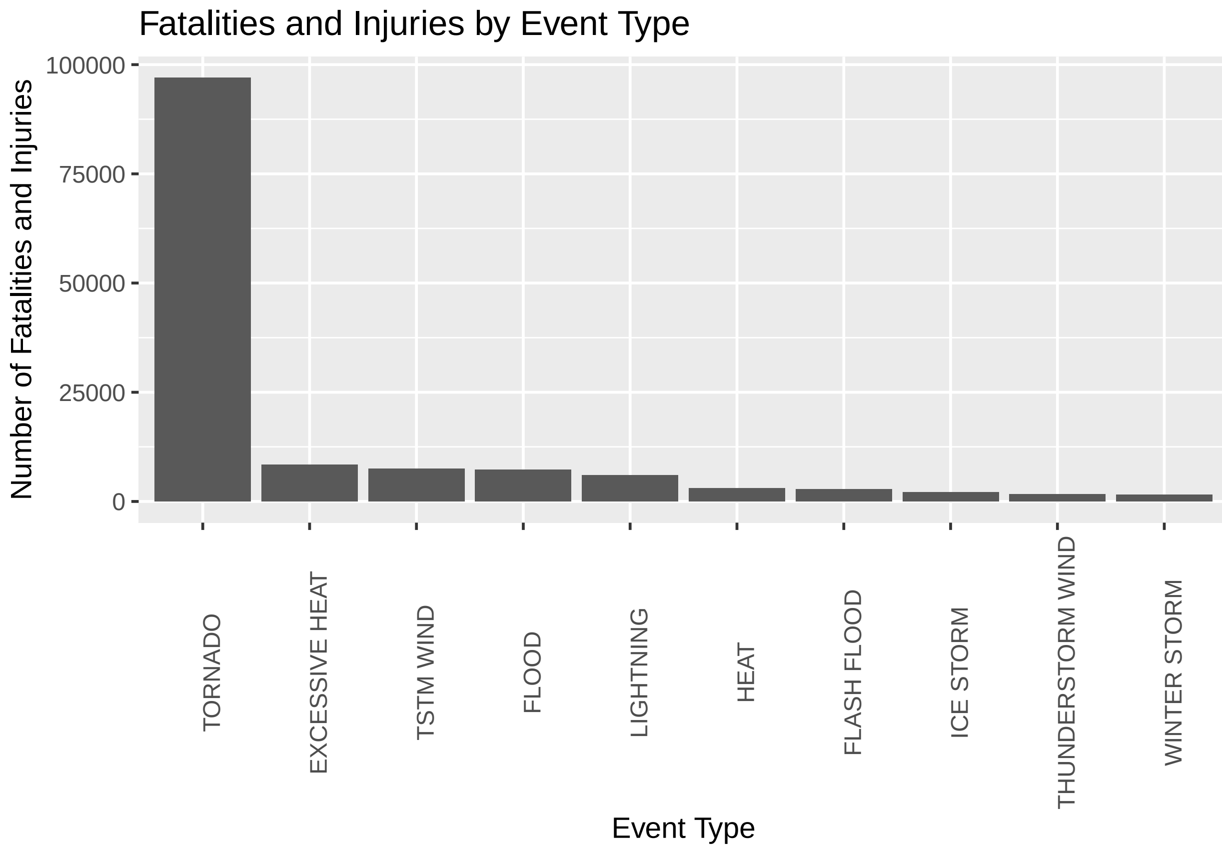
Next, when we see the data arranged by injuries, tornadoes again are first by a big margin.

```
inj_data
```

```
## # A tibble: 985 x 2
##    EVTYPE            tot_inj
##    <fct>               <dbl>
##  1 TORNADO             91346
##  2 TSTM WIND            6957
##  3 FLOOD                6789
##  4 EXCESSIVE HEAT       6525
##  5 LIGHTNING            5230
##  6 HEAT                 2100
##  7 ICE STORM            1975
##  8 FLASH FLOOD          1777
##  9 THUNDERSTORM WIND    1488
## 10 HAIL                 1361
## # ... with 975 more rows
```

Graph of top 10 events for fatalities plus injuries:

```
g <- ggplot(q1_data[1:10,], aes(reorder(EVTYPE, -total), total)) + geom_bar(stat="identity") +  theme(a
g
```



From this, we can see that tornadoes cause the most damage with respect to population health; they cause the most fatalities and most injuries.

Question 2

First, let's look at the top 10 events for property damage:

```
prop_data
```

```
## # A tibble: 985 x 2
##    EVTYPE             TotalPropertyDamage
##    <fct>                            <dbl>
##  1 FLOOD                     144657709800
##  2 HURRICANE/TYPHOON          69305840000
##  3 TORNADO                    56937160991
##  4 STORM SURGE                43323536000
##  5 FLASH FLOOD                16140812087.
##  6 HAIL                       15732266870
##  7 HURRICANE                  11868319010
##  8 TROPICAL STORM              7703890550
##  9 WINTER STORM                6688497250
## 10 HIGH WIND                   5270046260
## # ... with 975 more rows
```

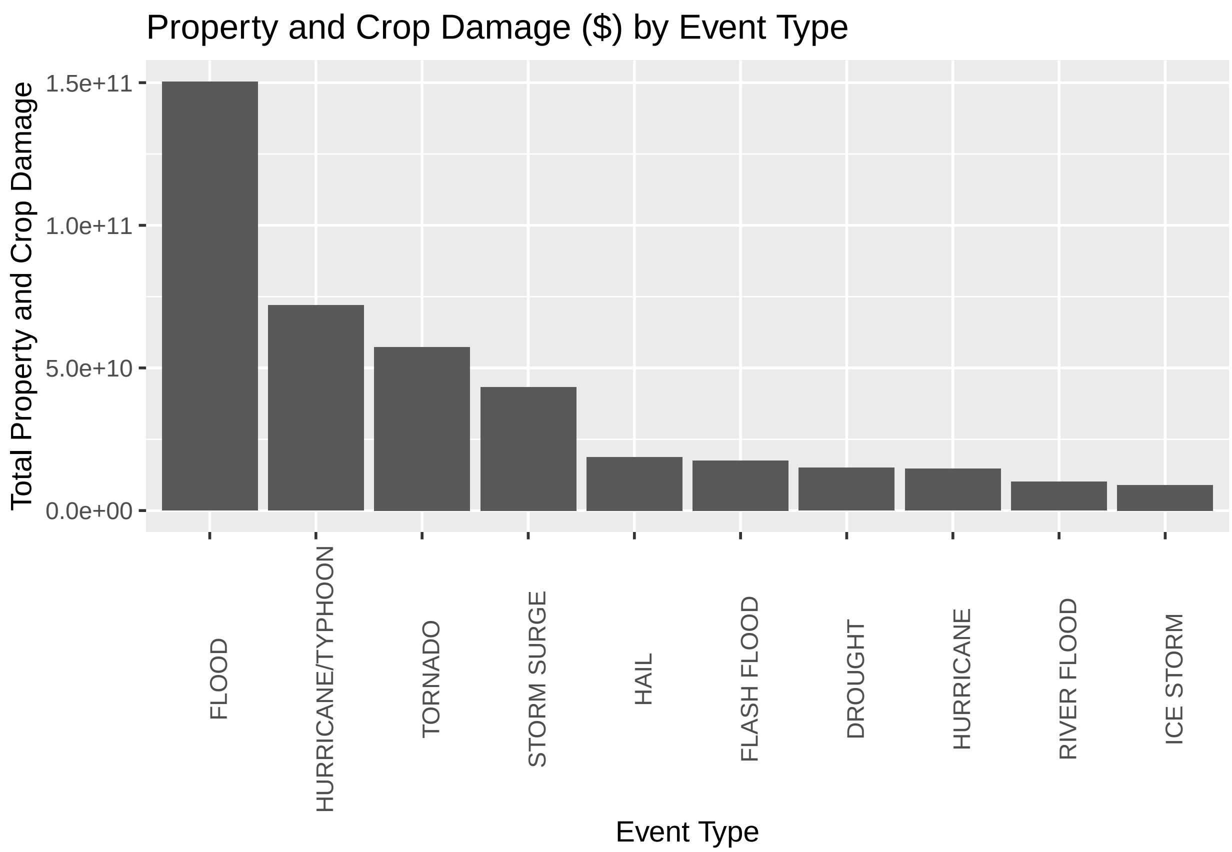
Then, for crop damage:

```r
crop_data
```

```
## # A tibble: 985 x 2
##    EVTYPE            TotalCropDamage
##    <fct>                       <dbl>
##  1 DROUGHT               13972566000
##  2 FLOOD                  5661968450
##  3 RIVER FLOOD            5029459000
##  4 ICE STORM              5022113500
##  5 HAIL                   3025954450
##  6 HURRICANE              2741910000
##  7 HURRICANE/TYPHOON      2607872800
##  8 FLASH FLOOD            1421317100
##  9 EXTREME COLD           1292973000
## 10 FROST/FREEZE           1094086000
## # ... with 975 more rows
```

Finally, we'll graph the top 10 events for the total property plus crop damage:

```r
g <- ggplot(q2_data[1:10,], aes(reorder(EVTYPE, -Total), Total)) + geom_bar(stat="identity") +  theme(a
g
```



From this, we can see that floods cause the most economic damage; they cause the most damage for both property and crops, and so also in total.