

Reproducible Research Project 2

We want to answer two questions using the NOAA data. - Across the United States, which types of events (as indicated in the `EVTTYPE` variable) are most harmful with respect to population health? - Across the United States, which types of events have the greatest economic consequences?

Pulling in the NOAA data file

```
rm(list=ls())
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.2.4
```

```
library(ggplot2)
```

```
setwd("/Users/marimuraki/Dropbox/Mari/courses/Coursera/NOAA")
```

```
url    <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
```

```
zip    <- "./data/repdata-data-StormData.csv.bz2"
```

```
file   <- "./data/repdata-data-StormData.csv"
```

```
if (!file.exists(zip)) {  
  download.file(url,  
                destfile=zip)  
}
```

```
if (!file.exists(file)) {  
  unzip(zip,  
        exdir="./data")  
  file.remove(zip)  
}
```

```
data <- read.csv(file,  
                 sep="," ,  
                 na.string = "NA",  
                 header=TRUE)
```

Subsetting to key variables

We are interested in analyzing the events that impact population health and economy. We will subset

the data to the relevant variables below.

- EVTYPE: Event Type (e.g. tornado, flood, etc.)
- FATALITIES: Number of fatalities
- INJURIES: Number of injuries
- PROPDMG: Property damage estimates, entered as actual dollar amounts
- PROPDMGEXP: Alphabetic Codes to signify magnitude “K” for thousands, “M” for millions, and “B” for billions)
- CROPDGMG: Crop damage estimates, entered as actual dollar amounts
- CROPDGMGEXP: Alphabetic Codes to signify magnitude “K” for thousands, “M” for millions, and “B” for billions)

```
sublist <- c("EVTYPE", "FATALITIES", "INJURIES", "PROPDMG", "PROPDMGEXP", "CROPDGMG", "CROPDGMGEXP")
stormdata <- subset(data, select=sublist)
```

Cleaning up main EVTYPE categories e.g., misspellings

The NOAA data file is quite messy. The data needs to be cleaned in order for accurate analyses. The variables we will need to clean for our analyses are:

- EVTYPE
- We need to clean up misspellings, distinct inputs to be consolidated, etc. For example, “tstm” re-coded with / as “thunderstorm”.
- CROPDGMG & CROPDGMGEXP; PROPDMG & PROPDGMGEXP
- We need to convert {CROPDGMG, PROPDMG} to actual dollar values using the multipliers {CROPDGMGEXP, PROPDGMGEXP}

Cleaning EVTYPE values

```
stormdata$EVTYPE <- tolower(stormdata$EVTYPE)
stormdata$EVTYPE[grepl("blizzard", stormdata$EVTYPE, ignore.case = TRUE)]
  <- "blizzard"
stormdata$EVTYPE[grepl("cold", stormdata$EVTYPE, ignore.case = TRUE)]
  <- "cold"
stormdata$EVTYPE[grepl("fire", stormdata$EVTYPE, ignore.case = TRUE)]
  <- "fire"
stormdata$EVTYPE[grepl("flood", stormdata$EVTYPE, ignore.case = TRUE)]
  <- "flood"
stormdata$EVTYPE[grepl("hail", stormdata$EVTYPE, ignore.case = TRUE)]
  <- "hail"
stormdata$EVTYPE[grepl("heat", stormdata$EVTYPE, ignore.case = TRUE)]
  <- "heat"
stormdata$EVTYPE[grepl("high surf", stormdata$EVTYPE, ignore.case = TRUE)]
  <- "high surf"
stormdata$EVTYPE[grepl("hurricane", stormdata$EVTYPE, ignore.case = TRUE)]
  <- "hurricane"
stormdata$EVTYPE[grepl("lightn", stormdata$EVTYPE, ignore.case = TRUE)]
  <- "lightning"
stormdata$EVTYPE[grepl("mud.*slide", stormdata$EVTYPE, ignore.case = TRUE)]
  <- "mudslide"
stormdata$EVTYPE[grepl("rain", stormdata$EVTYPE, ignore.case = TRUE)]
  <- "rain"
stormdata$EVTYPE[grepl("precip", stormdata$EVTYPE, ignore.case = TRUE)]
  <- "rain"
stormdata$EVTYPE[grepl("rip current", stormdata$EVTYPE, ignore.case = TRUE)]
  <- "rip current"
stormdata$EVTYPE[grepl("snow", stormdata$EVTYPE, ignore.case = TRUE)]
  <- "snow"
stormdata$EVTYPE[grepl("storm surge", stormdata$EVTYPE, ignore.case = TRUE)]
  <- "storm surge"
stormdata$EVTYPE[grepl("thun.*orm", stormdata$EVTYPE, ignore.case = TRUE)]
  <- "thunderstorm"
stormdata$EVTYPE[grepl("tstm", stormdata$EVTYPE, ignore.case = TRUE)]
  <- "thunderstorm"
stormdata$EVTYPE[grepl("tornad", stormdata$EVTYPE, ignore.case = TRUE)]
  <- "tornado"
stormdata$EVTYPE[grepl("tropical.*storm", stormdata$EVTYPE, ignore.case = TRUE)]
] <- "tropical storm"
stormdata$EVTYPE[grepl("wind", stormdata$EVTYPE, ignore.case = TRUE)]
  <- "wind"
stormdata$EVTYPE[grepl("winter.*mix", stormdata$EVTYPE, ignore.case = TRUE)]
  <- "winter mix"
stormdata$EVTYPE[grepl("winter storm", stormdata$EVTYPE, ignore.case = TRUE)]
  <- "winter storm"
stormdata$EVTYPE[grepl("volcanic", stormdata$EVTYPE, ignore.case = TRUE)]
```

```
<- "volcanic"
```

Converting CROPDMG & PROPDMG

```
stormdata$CROPDMGEXP <- tolower(stormdata$CROPDMGEXP)
stormdata$PROPDMGEXP <- tolower(stormdata$PROPDMGEXP)
unique(stormdata$CROPDMGEXP)
```

```
## [1] "" "m" "k" "b" "?" "0" "2"
```

```
unique(stormdata$PROPDMGEXP)
```

```
## [1] "k" "m" "" "b" "+" "0" "5" "6" "?" "4" "2" "3" "h" "7" "-" "1" "8"
```

```
stormdata$CROPDMGEXP[stormdata$CROPDMGEXP %in% c("+","0","?","-","")] <- 0
stormdata$CROPDMGEXP[stormdata$CROPDMGEXP == "k"] <- 3
stormdata$CROPDMGEXP[stormdata$CROPDMGEXP == "m"] <- 6
stormdata$CROPDMGEXP[stormdata$CROPDMGEXP == "h"] <- 2
stormdata$CROPDMGEXP[stormdata$CROPDMGEXP == "b"] <- 9
unique(stormdata$CROPDMGEXP)
```

```
## [1] "0" "6" "3" "9" "2"
```

```
stormdata$PROPDMGEXP[stormdata$PROPDMGEXP %in% c("+","0","?","-","")] <- 0
stormdata$PROPDMGEXP[stormdata$PROPDMGEXP == "k"] <- 3
stormdata$PROPDMGEXP[stormdata$PROPDMGEXP == "m"] <- 6
stormdata$PROPDMGEXP[stormdata$PROPDMGEXP == "h"] <- 2
stormdata$PROPDMGEXP[stormdata$PROPDMGEXP == "b"] <- 9
unique(stormdata$PROPDMGEXP)
```

```
## [1] "3" "6" "0" "9" "5" "4" "2" "7" "1" "8"
```

```
stormdata$CROPDMG_dollars <- stormdata$CROPDMG * 10^as.numeric(stormdata$CROPDMGEXP)
stormdata$PROPDMG_dollars <- stormdata$PROPDMG * 10^as.numeric(stormdata$PROPDMGEXP)
```

Analysis & Results

Question 1: Across the United States, which types of events (as indicated in

the EVTYPE variable) are most harmful with respect to population health?

We will define events “harmful with respect to population health” to be the combination of fatalities and injuries.

```
harm_pophealth <- aggregate(list (harm = stormdata$FATALITIES + stormdata$INJURIES),
                             list (EVTYPE = stormdata$EVTYPE),
                             sum)
harm_pophealth <- harm_pophealth[with (harm_pophealth, order (harm, decreasing=TRUE)),]
head(harm_pophealth)
```

```
##           EVTYPE  harm
## 249      tornado 97043
## 85         heat 12362
## 248 thunderstorm 10172
## 62         flood 10129
## 126    lightning  6049
## 297         wind  2379
```

Let us identify the top 10 events harmful to population health. We see that tornados account by far the most to total fatalities and injuries. Tornados are 8x more than the second highest contributor, heat.

```
top_harm_pophealth <- harm_pophealth[order(-harm_pophealth$harm),][1:10,]
list(top_harm_pophealth)
```

```
## [[1]]
##           EVTYPE  harm
## 249      tornado 97043
## 85         heat 12362
## 248 thunderstorm 10172
## 62         flood 10129
## 126    lightning  6049
## 297         wind  2379
## 117     ice storm  2064
## 59         fire  1698
## 299 winter storm  1554
## 82         hail  1512
```

Visualizing these cumulative event counts over time, we can see how high tornado events are compared to the other events affecting population health.

```
png(file="plot1_harm_pophealth.png")
ggplot(top_harm_pophealth, aes(x = reorder(EVTYPE, -harm), y = harm)) +
  geom_bar(stat = "identity", aes(fill = harm), position = "dodge") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  xlab("Event Type") +
  ylab("Total Events") +
  ggtitle("Harmful events (fatalities + injuries) to population health in USA")
dev.off()
```

```
## quartz_off_screen
##                               2
```

Q2: Across the United States, which types of events have the greatest economic consequences?

We will define events having the “greatest economic consequences” to be the combination of crop and property damage.

```
harm_econ <- aggregate(list (harm = stormdata$CROPDMG_dollars + stormdata$PROPD
MG_dollars),
                      list (EVTYPE = stormdata$EVTYPE),
                      sum)
harm_econ <- harm_econ[with (harm_econ, order (harm, decreasing=TRUE)),]
head(harm_econ)
```

```
##           EVTYPE           harm
## 62           flood 180592324935
## 106    hurricane  90271472810
## 249     tornado  57418279346
## 180 storm surge  47965579000
## 82           hail  20737204410
## 33     drought  15018672000
```

Let us identify the top 10 events harmful to the economy. Floods are the top contributors to economic harm, followed by hurricane and tornadoes.

```
top_harm_econ <- harm_econ[order(-harm_econ$harm),][1:10,]
list(top_harm_econ)
```

```
## [[1]]
##           EVTYPE           harm
## 62           flood 180592324935
## 106          hurricane 90271472810
## 249           tornado 57418279346
## 180      storm surge 47965579000
## 82            hail 20737204410
## 33            drought 15018672000
## 248   thunderstorm 12345977914
## 117         ice storm 8967041360
## 59            fire 8904910130
## 252 tropical storm 8409286550
```

Visualizing these cumulative event counts over time, we see that floods, hurricanes and tornadoes are the top contributors to economically harmful events — with floods accounting for 2x that of hurricanes.

```
png(file="plot1_harm_econ.png")
ggplot(top_harm_econ, aes(x = reorder(EVTYPE, -harm), y = harm)) +
  geom_bar(stat = "identity", aes(fill = harm), position = "dodge") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  xlab("Event Type") +
  ylab("Total Events") +
  ggtitle("Harmful economic events (crop + property damage) to US economy")
dev.off()
```

```
## quartz_off_screen
##                2
```

Future Analysis

Future analysis will include: - Events by year: Rather than aggregating all events over time, we can examine how events change over time. - Break out events: Rather than combining, for example, fatalities + injuries, we can examine how each contribute.