Coursera Class: Reproducible Reasearch Week4 - Peer Review Assignment

April L

January 20, 2018

Introduction

Storms and other sever wether events can cause both public health and economic problems for communities and munipalitie. Many server 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 US National Oceanic and Atmospheric Administration's (NOAA) storm database. This database tracks characteristic 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 thsi assignment is to explore the NOAA storm database and answer some basic questiosn 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, and other summaries. You may use any R package you want to support your analysis.

The 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 consequencs?

Property Damage Crop Damage Flood-Related Damage, each WFO report flood damage in their County Warning Area (CWA) WFO is a person

Data Processing

** Read Data **

First, load the data from the file into R.

After a glance at the data, I found that there are inconsistency in the EVTYPE. In the column, there are mis-spelling, different meaning, and observations which could be out of scope.

Using the suggestion in : https://rstudio-pubs-static.s3.amazonaws.com/58957_37b6723ee52b455990e149edde45e5b6.html I calculate the property damage value.

```r
temp <- tempfile()

if (!file.exists(temp)) {
    download.file("https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2", temp)
}

data <- read.csv(temp)
```

explore & clean EVTYPE data

1. Remove Leading/Trailing Whitespace

```
# remove leading & trailing whitespace
data$EVTYPE_c <- trimws(data$EVTYPE, which = c('both'))

#assume the "Monthly" & "Summary" are not the special events
cond <- "Monthly|Summary"
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data <- data[-(c),]


# explore ETTYPE
a <- sort(unique(data$EVTYPE))
```

```
#create a new column for cleaned EVTYPE

cond <- "^AVALAN.*$"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data$EVTYPE_c[c] <- "Avalanche"

cond <- "^BLIZZARD.*$"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data$EVTYPE_c[c] <- "Blizzard"

cond <- "Dust"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data$EVTYPE_c[c] <- "Dust Storm"

cond <- "Smoke"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data$EVTYPE_c[c] <- "Dense Smoke"

cond <- "Funnel.*$"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data$EVTYPE_c[c] <- "Funnel Cloud"

cond <- "Surf"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data$EVTYPE_c[c] <- "High Surf"

cond <- "^WINTER S.*$"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data$EVTYPE_c[c] <- "Winter Storm"

cond <- "^WINTER W.*$"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data$EVTYPE_c[c] <- "Winter Weather"
```

```
cond <- "^Wild.*$"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data$EVTYPE_c[c] <- "Wildfire"

cond <- ".*SPOUT.*$"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data$EVTYPE_c[c] <- "Waterspout"

cond <- "^VO.*$"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data$EVTYPE_c[c] <- "Volcanic Ash"

cond <- "Tropical Storm.*$"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data$EVTYPE_c[c] <- "Tropical Storm"

cond <- "^TORNA?Da?O.*$"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data$EVTYPE_c[c] <- "Tornado"

cond <- "Sleet"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data$EVTYPE_c[c] <- "Sleet"

cond <- "Heat"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data$EVTYPE_c[c] <- "Heat"

cond <- "COLD"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data$EVTYPE_c[c] <- "Cold/Wind Chill"

cond <- "Rain"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data$EVTYPE_c[c] <- "Heavy Rain"

cond <- "Heavy Snow"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data$EVTYPE_c[c] <- "Heavy Snow"

cond <- "Snow"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
```

```r
data$EVTYPE_c[c] <- "Heavy Snow"

cond <- "High *Wind"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data$EVTYPE_c[c] <- "High Wind"

cond <- "^LIGHTN?ING.*$"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data$EVTYPE_c[c] <- "Lightning"
c <- which(data$EVTYPE_c == 'LIGNTNING')
data$EVTYPE_c[c] <- "Lightning"

cond <- "(?|(Hurricane)|(TYPHOON))"
b <- unique(data$EVTYPE_c[grepl(cond , data$EVTYPE_c, ignore.case = TRUE, perl=T)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)
data$EVTYPE_c[c] <- "Hurricane (Typhoon)"

cond <- "(?|DROUGHT|Dry)"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)
data$EVTYPE_c[c] <- "Drought"

###

cond <- "^THUNDER.*$"
cond <- "(?<!Marine )thu.*"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)
data$EVTYPE_c[c] <- "Thunderstorm Wind"

cond <- "^TSTM.*$"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data$EVTYPE_c[c] <- "Thunderstorm Wind"

###

cond <- "(?<!Marine )(?|(Strong)|(Gusty)) WIND"   # Not begin with Marine, Preceed with Strong or Gusty
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)
data$EVTYPE_c[c] <- "Strong Wind"

cond <- ".?(?<!Marine )Hail.*$"    #Not begin with Marine
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE, perl=TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)
data$EVTYPE_c[c] <- "Hail"

c <- which(data$EVTYPE_c == 'MARINE TSTM WIND')
data$EVTYPE_c[c] <- "MARINE THUNDERSTORM WIND"

cond <- ".?Flash.*$"
```

```r
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE)
data$EVTYPE_c[c] <- "Flash Flood"

cond <- "(?|(Coast)|(Beach)|(CSTL))"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)
data$EVTYPE_c[c] <- "Coastal Flood"

cond <- "(?<!Coastal |Flash )flood"   #not begin with Coastal or Flash
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)
data$EVTYPE_c[c] <- "Flood"

cond <- "Extreme (?! WET)"   #not follow with WET
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)
data$EVTYPE_c[c] <- "Extreme Cold/Wind Chill"

cond <- "Freez(?!ing FOG)"   #not follow with ing FOG
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)
data$EVTYPE_c[c] <- "Frost/Freeze"

cond <- "Freez(?=ing FOG)"   #not follow with ing FOG
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)
data$EVTYPE_c[c] <- "Freezing Fog"

cond <- "Ice"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)
data$EVTYPE_c[c] <- "Ice Storm"


cond <- "Lake.Effect"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)
data$EVTYPE_c[c] <- "Lake-Effect Snow"

cond <- "^Wint...(?!Storm|Weather)"
b <- unique(data$EVTYPE_c[grepl(cond, data$EVTYPE_c, ignore.case = TRUE, perl=TRUE)])
c <- grep(cond, data$EVTYPE_c, ignore.case = TRUE, perl=T)
data$EVTYPE_c[c] <- "Wintry Mix"
```

```r
a <- sort(unique(data$EVTYPE_c))

# take a look of number of observations cleaned **
file_path = "~/workspace/AL-DataScienceCoursera/Reproducible Research"
analysis_event = read.csv(file.path(file_path,'StormEventTable.csv'))

temp = data[data$EVTYPE_c %in% analysis_event$Event,]

cleaned = round(nrow(temp)/nrow(data)*100,2)
```

After cleaning the EVTYPE values, 97.63% observations are cleaned / avialable for analysis.

** process property damage value **

```
summary(temp$PROPDMGEXP)
```

```
##             -      ?      +      0      1      2      3      4      5 
## 453839      1      8      5    215     25     13      4      4     28 
##      6      7      8      B      h      H      K      m      M 
##      4      5      1     37      1      6 415382      7  11215 
```

```
temp$PROPDMGEXP <- as.character(temp$PROPDMGEXP)

a <- which((temp$PROPDMGEXP %in% seq(2,8)))
temp[a,'PROPDamage'] <- temp[a,'PROPDMG'] * as.numeric(temp[a,'PROPDMGEXP'])

a <- which((temp$PROPDMGEXP == 'H') | (temp$PROPDMGEXP == 'h'))
temp[a,'PROPDamage'] = temp[a,'PROPDMG'] * 100

a <- which((temp$PROPDMGEXP == 'K') | (temp$PROPDMGEXP == 'k'))
temp[a,'PROPDamage'] = temp[a,'PROPDMG'] * 1000

a <- which((temp$PROPDMGEXP == 'M') | (temp$PROPDMGEXP == 'm'))
temp[a,'PROPDamage'] = temp[a,'PROPDMG'] * 1000000

a <- which((temp$PROPDMGEXP == 'B') | (temp$PROPDMGEXP == 'b'))
temp[a,'PROPDamage'] = temp[a,'PROPDMG'] * 1000000000

a <- which((temp$PROPDMGEXP == '+') | (temp$PROPDMGEXP == '1'))
temp[a,'PROPDamage'] = temp[a,'PROPDMG']

a <- which((temp$PROPDMGEXP == '0') )
temp[a,'PROPDamage'] = temp[a,'PROPDMG'] * 10

a <- which((temp$PROPDMGEXP == '')
           | (temp$PROPDMGEXP == '?')
           | (temp$PROPDMGEXP == '-'))
temp[a,'PROPDamage'] = temp[a,'PROPDMG'] * 0
```

Results

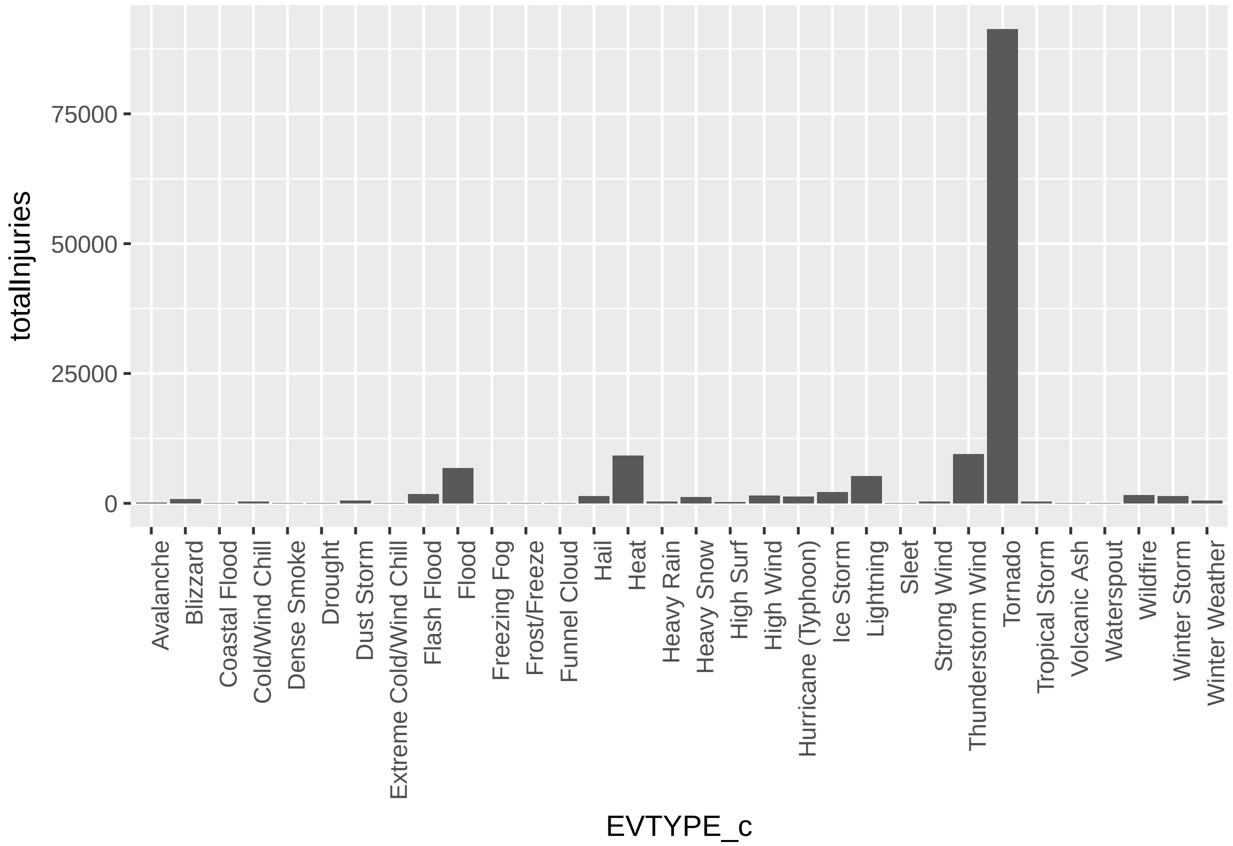
** Question 1 ** 1. Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health?

```
a <- temp %>%
     group_by(EVTYPE_c) %>%
     summarize(totalInjuries = sum(INJURIES))

event <- a[which.max(a$totalInjuries),'EVTYPE_c']



ggplot(data=a, aes(x=EVTYPE_c, y=totalInjuries)) +
    geom_bar(stat="identity") +
    theme(axis.text.x = element_text(angle=90, hjust=1))
```



Question 1 Result: As show in the plot, across the United States, Tornado is tne most harmful with respect to population health.

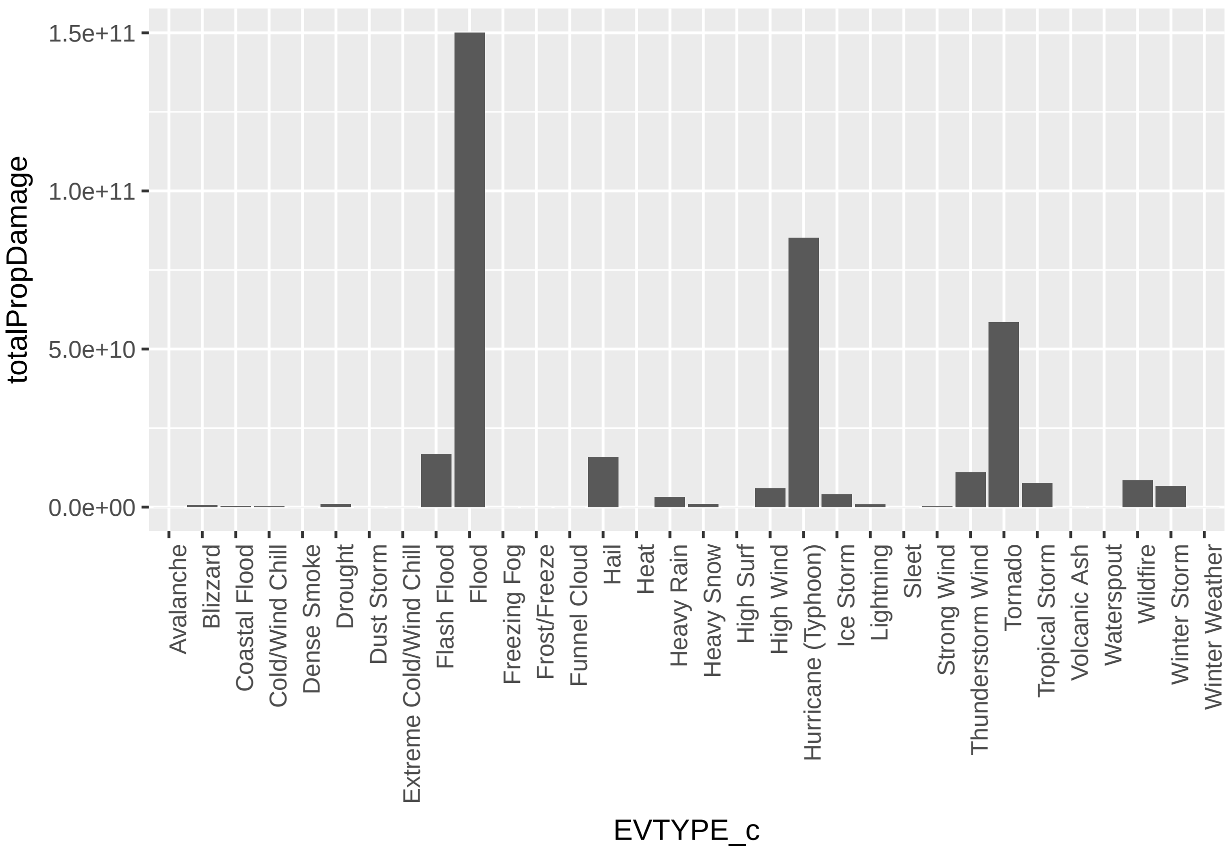
** Question 2 ** Across the United States, which types of events have the greatest economic consequences?

```
a <- temp %>%
     group_by(EVTYPE_c) %>%
     summarize(totalPropDamage = sum(PROPDamage))

event <- a[which.max(a$totalPropDamage),'EVTYPE_c']

library(ggplot2)

ggplot(data=a, aes(x=EVTYPE_c, y=totalPropDamage)) +
    geom_bar(stat="identity") +
    theme(axis.text.x = element_text(angle=90, hjust=1))
```



Question 2 Result: As show in the plot, across the United States, Flood have the greatest economic consequeces. “‘

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.